Predictability of Anomalous Storm Tracks from Seasonal to Decadal Scales

Gilbert P. Compo and Prashant D. Sardeshmukh

NOAA-CIRES Climate Diagnostics Center

University of Colorado at Boulder

Submitted to the Journal of Climate

September 10, 2003

Corresponding author: Dr. Gilbert P. Compo, NOAA-CIRES Climate Diagnostics Center R/CDC1, 325 Broadway, Boulder, CO 80305-3328. (303) 497-6115. compo@colorado.edu

Abstract

This paper is concerned with estimating the predictable variation of extratropical daily weather statistics ("stormtracks") associated with sea surface temperature (SST) changes on interannual to interdecadal scales, and its magnitude relative to the unpredictable noise. The SST-forced stormtrack signal in each winter in 1950-99 is defined as the mean stormtrack anomaly obtained in an ensemble of atmospheric general circulation model (GCM) integrations with prescribed observed SSTs. Two sets of relatively small (9- to 13-member) ensembles available from two modeling centers (NCAR and NCEP), with anomalous SSTs prescribed either globally or in the tropics alone, are used. Since the stormtrack signals cannot be derived directly from the archived GCM output, they are diagnosed from the SST-forced winter-mean 200 mb height signals using an empirical linear stormtrack model (STM). For two particular winters (the El Nino of JFM 1987 and the La Nina of JFM 1989), the stormtrack signals and noise are estimated directly, and more accurately, from additional large (60-member) ensemble runs of the NCEP GCM. The linear STM is shown to be remarkably successful at capturing the GCM's stormtrack signal in these two winters, and is thus suitable for estimating the signal in other winters.

The principal conclusions from this analysis are as follows. A predictable SST-forced stormtrack signal exists in many winters, but its strength and pattern can change substantially from winter to winter. The correlation of the SST-forced and observed stormtrack anomalies is high enough in the Pacific-North American (PNA) sector to be of practical use. Most of the SST-forced signal is associated with tropical Pacific SST forcing; the central Pacific (Nino-4) is somewhat more important than the eastern Pacific (Nino-3) in this regard. Variations of the pattern correlation of the SST-forced and observed stormtrack anomaly fields from winter to winter, and among 5-winter averages, are generally consistent with variations of the signal strength, and to that extent are identifiable *a priori*. Larger pattern correlations for the 5-winter averages in the second half of the 50-yr record, and also the 50-yr stormtrack trend, are consistent with the stronger ENSO SST forcing in the second half. *None of these conclusions, however, apply in the Euro-Atlantic*

sector, where the correlations of the SST-forced and observed stormtrack anomalies are found to be much smaller. Given also that they are inconsistent with the estimated signal to noise ratios, substantial GCM error in representing the response in this region to tropical SST forcing, rather than intrinsically low Euro-Atlantic stormtrack predictability, is argued to be behind these lower correlations.

1. Introduction.

It is well known that the statistics of extratropical daily weather ("stormtracks") averaged over individual winter seasons, decades, and even longer intervals are not constant but vary substantially from one interval to the next. These variations have a random part associated with sampling fluctuations and a potentially predictable part associated with slow changes in atmospheric boundary conditions and atmospheric composition. This paper addresses the problem of estimating the predictable signal associated specifically with sea surface temperature (SST) changes, and its magnitude relative to the random noise. This signal to noise ratio has a simple relationship to the potential skill of stormtrack anomaly forecasts, and is thus a useful measure of stormtrack predictability.

Many studies have assessed the predictability of atmospheric variations associated with anomalous SSTs, especially those associated with the ENSO phenomenon. They have focused mostly on seasonal to multiyear averages of a few select variables such as geopotential height, surface temperature, and precipitation (e.g. Kumar et al. 1996; Chen and van den Dool 1997; Brankovic and Palmer 1997, 2000; Rowell 1998; Anderson et al. 1999; Koster et al. 2000; Graham et al. 2000; Shukla et al. 2000; Peng et al. 2000, 2002; Zwiers et al. 2000 and references therein; Kumar et al. 2003). As demonstrated, however, in several recent studies using large AGCM ensembles (Sardeshmukh et al. 2000, Schubert et al. 2001, Compo et al. 2001), ENSO-related SST anomalies affect more than the mean circulation around the globe. They also alter the variability — indeed the entire probability distribution — of atmospheric variables from daily to seasonal scales. The predictability of the second and higher moments of the distribution has thus far received almost no attention.

There is substantial observational evidence of an ENSO effect on northern hemispheric stormtracks, extending eastward from the central north Pacific across north America and the Atlantic to Europe (e.g. Fraedrich 1990, 1994, Fraedrich and Muller 1992, Hoerling and Ting 1994, Straus and Shukla 1997, May and Bengtsson 1998, Matthews and Kiladis 1999, Smith and Sardeshmukh 2000, Sardeshmukh et al. 2000, Carillo et al. 2000, Compo et al. 2001). The

observations also suggest a somewhat different effect for El Nino and La Nina forcing, thus hinting that it may vary from ENSO case to case, with obvious implications for stormtrack predictability. The limited observational record, however, compromises estimating such case-dependent and/or nonlinear signals with statistical significance, and it also compromises estimating the stormtrack noise essential for assessing predictability. A similar remark applies to most previous assessments of these effects made using small AGCM ensembles. To remedy this situation, Compo et al. (2001) examined much larger 60-member ensembles of seasonal NCEP AGCM integrations with prescribed observed global SSTs for one El Nino (JFM 1987) and one La Nina (JFM 1989) case, and were able to conclude with much greater confidence that the SST-forced stormtrack signal may indeed vary substantially from case to case, especially over the North Atlantic and Europe. They could also demonstrate a statistically significant stormtrack signal in many regions not usually associated with an ENSO effect.

Demonstrating the existence of a signal is, of course, not the same as demonstrating its predictability or usefulness. What matters for predictability, and usefulness, is the size of the signal relative to the noise. Compo et al. did not consider this question explicitly, our principal concern here.

The signal and noise for any quantity may be estimated from ensemble integrations as the ensemble mean anomaly and ensemble spread, respectively. The upper right panel of Fig. 1 shows their ratio S for the SST-forced stormtrack anomalies in JFM 1987 estimated from the 60-member NCEP AGCM ensembles considered by Compo et al. To generate this plot, the winter (JFM) mean stormtrack anomaly was defined for each ensemble member at each geographical location as the deviation of the winter-mean 2-to-7 day bandpass variance of 500 mb vertical velocity (omega) from the ensemble-mean bandpass variance obtained in a 90-winter ensemble with prescribed climatological JFM SSTs. The stormtrack signal associated with JFM 1987 SSTs was then obtained as the ensemble-mean of the 60 stormtrack anomaly values, and the noise as the rms deviation of the 60 values from this mean. Figure 1 also shows the S values for the SST-forced winter-mean 500 mb omega and precipitation anomalies obtained in a similar manner. As

suspected, the stormtrack S values are modest, but comparable in magnitude to the S values for 500 mb omega and precipitation. Indeed the figure strongly suggests that the predictability of winter-mean precipitation is as much tied to the predictability of the winter-mean 500 mb omega stormtrack as to that of the winter-mean 500 mb omega. Mainly for this reason, we will restrict ourselves to the predictability of the "500 mb omega stormtrack", as opposed to many other interesting measures of synoptic variability.

As discussed by Sardeshmukh et al. (2000) and many others (see Appendix), the signal-to-noise ratio S for any quantity (local anomalies or anomaly fields; first or higher moments) has a simple monotonic relationship with the expected correlation skill ρ_n of n-member ensemble-mean forecasts made using a "perfect" model. Specifically,

$$\rho_n = S^2 / [(S^2 + 1)(S^2 + n^{-1})]^{1/2} . \qquad (1)$$

The thin curves in Figure 2 illustrate this relationship for a few values of n. S is thus a useful measure of predictability. The outermost ρ_∞ curve shows how predictability is limited if S is small; this limitation cannot be overcome even using infinite-member ensembles of a perfect model. The expected skill ρ_n using n-member ensembles is lower than ρ_∞, and model error (see Appendix) gives rise to even lower actual skill ρ.

The modest estimates of S for the omega stormtrack in Fig. 1 therefore imply only modest stormtrack predictability associated with SST changes. It is important to keep in mind, however, that these estimates are for one specific winter case using one particular AGCM. It is unclear to what extent they are affected by the specifics of that case and/or model error. Comparing the GCM's ensemble-mean predicted stormtrack anomaly with the observed anomaly in JFM 1987 does not settle the issue, because the GCM's prediction is only the *expected* anomaly. The prediction problem is inherently probabilistic, so the reliability of model-generated predictability estimates can only be assessed by examining prediction skill over a large number of cases. To this end, one should ideally generate similar estimates of S for all the past 50+ winters for which observational upper-air verification data are available, using similar large 60-member ensemble runs of several other AGCMs to estimate the SST-forced stormtrack signal and noise in each

winter. One could then make a scatter plot of the actual ρ against S, and determine to what extent the points fall along the ρ_{60} curve in Fig. 2. To that extent, one would feel confident in the stormtrack predictability estimates derived from those AGCMs. One could then consider the histogram of S values along the abscissa of Fig. 2, and use it to estimate, from the ρ_{60} curve, the mean expected skill, i.e. the mean predictability, as well as the range of the predictable skill variation around this mean predictability. On the other hand, if the scatter points fall well below the ρ_{60} curve, then one would have to conclude that model error was compromising the predictability estimates. To our knowledge it is not yet possible to make such scatter plots, because the necessary large ensemble runs with archived daily output have not yet been made at modeling centers. We have nevertheless attempted in the final figure of this paper (Fig. 13) to make such a plot using smaller 12-member ensemble runs of two AGCMs for 1950-99. It provides perhaps the best overall assessment to date of northern hemispheric stormtrack predictability, although it will be clear that much else remains to be done.

The question of what ensemble sizes to use to make such assessments is particularly important, but is often glossed over. Are 12 members enough? One might think so from Fig. 2, given that the ρ_{12} curve is "close enough" to the ρ_{∞} curve. The challenge in correctly estimating predictability, however, lies in correctly estimating S. The thickened portions of the curves in Fig. 2 illustrate the 1-standard deviation uncertainty in estimates of $S = 0.7$ using n members. The large error bars on S for $n = 12$ (and even $n = 25$) translate into correspondingly large errors bars on the expected skill, i.e the predictability estimates. Clearly, n must be sufficiently large that the sampling uncertainty in S is smaller than the actual variation of S from case to case. Otherwise, one loses the ability to determine and exploit the case-dependent variations of predictability. Sampling errors in S can also make the expected skill appear spuriously inconsistent with the actual skill, and may thus cause them to be confused with model errors.

Our principal means of making a general assessment of wintertime stormtrack predictability in this paper will be to compare the SST-forced variations of the stormtracks over the last half-century with the observed variations. To this end, we will use two sets of relatively

small (9- to 13-member) ensemble runs for the last half century available from two modeling centers (NCAR and NCEP), with anomalous observed SSTs prescribed either globally or in the tropics alone, to estimate the SST-forced signal in each winter. Since the stormtrack signals cannot be obtained directly from the archived monthly GCM output, we will diagnose them from the winter-mean 200 mb height signals using an empirical linear stormtrack model (STM) developed specifically for this purpose. We will present the correlations ρ of these STM-diagnosed stormtrack signals with the observed stormtrack anomalies, both as maps of the correlation of the SST-forced and observed anomaly values over 50 winters at each gridpoint, and as 50-winter timeseries of the pattern correlation of the SST-forced and observed anomaly fields in each winter over the Pacific-North American (PNA) and North Atlantic-Europe (NATL-EUR) sectors.

In terms of our preceding discussion, we will thus use the actual skill ρ rather than directly estimating S to assess stormtrack predictability. It is clear from Fig. 2 that using ~12 member ensembles is inadequate for estimating predictability in individual winters at individual gridpoints. However, the error bars on 50-winter correlations at individual grid points are much smaller, and are also smaller in individual winters for pattern correlations over the relatively large PNA and NATL-EUR domains with several spatial degrees of freedom. This will justify our presenting the maps of local correlations and the timeseries of pattern correlations discussed above. The latter, in comparison with the timeseries of tropical SST indices, will help us assess the case-dependent variations of stormtrack predictability.

Our main interest in this paper is in extracting the predictable component of the stormtrack variations. The words "predictable" and "SST-forced" can be used almost interchangeably for interannual stormtrack variations, given the large influence of predictable interannual tropical SST variations in forcing them. Several studies have also shown substantial decadal stormtrack variability and trends over the last 50 to 100 years (e.g. Hurrell and van Loon 1997, WASA Group 1998, Graham and Diaz 2001, Chang and Fu 2002, 2003, Gulev et al. 2002, Harnik and Chang 2003). The decadal variations of the omega stormtrack, with its more direct

link to precipitation variations, have not been previously studied, and the degree to which they are SST-forced has also not been addressed. Acknowledging that the existence of an SST-forced component does not necessarily imply stormtrack predictability on decadal scales, we will nevertheless also present correlations of 5-winter averages of the SST-forced and observed stormtrack anomalies, and explore to what extent they are associated with anomalous 5-winter average tropical SSTs.

The paper is organized as follows. The data and model integrations are discussed in section 2. In section 3, the linear STM is developed and tested for its ability to reproduce the NCEP AGCM's (60-member) SST-forced stormtrack signals in JFM 1987 and 1989, given only the AGCM's ensemble-mean 200 mb height signals in those winters. In section 4, the STM is used to diagnose the SST-forced stormtrack signals in 1950-1999, given the NCEP and NCAR AGCMs' ensemble-mean 200 mb height responses to (a) observed Global SST forcing, and (b) observed Tropical SST forcing. The AGCMs' skill in simulating the observed stormtrack anomalies is then evaluated through the correlation measures discussed above. This average skill is compared with that expected from Eq (1) for the 1987 and 1989 events to illustrate the case-dependence of expected forecast skill. In section 5, the STM is used to address the important issue of decadal variations of stormtrack activity. In section 6, we show that the actual AGCM skill in predicting stormtracks is close to the expected skill over the Pacific-North American sector, but a substantial systematic error is present over the North Atlantic-European sector. A discussion and concluding remarks follow in section 7.

2. Data.

NCEP and NCAR AGCM simulation data, Hadley Centre SST data, and NCEP-NCAR reanalysis data are used in this study. The NCEP model used is the MRF9, identical to that used by Kumar et al. (1996), Chen and van den Dool (1997), Sardeshmukh et al. (2000), and Compo et al. (2001). The model has a spatial discretization of T40 in the horizontal (about 3^o lat by 3^o lon) and 18 sigma (normalized pressure) levels in the vertical. Kanamitsu et al. (1991) describe

the NCEP MRF model in detail. The several sets of MRF9 integrations used are listed in Table 1. The first set is the same as made by Sardeshmukh et al. 2000 and Compo et al. 2001. Large ensembles of seasonal integrations were made with observed monthly global climatological JFM SSTs (90 members) and observed monthly global SSTs for JFM 1987 (60 members) and JFM 1989 (60 members). The second set was made with observed monthly global climatological JFM SSTs specified in all locations except the Nino4 region (5^oN - 5^oS, 160^oE – 150^oW). As described in Table 1, we have made a total of 360 integrations of the MRF9 with SST anomalies in the Nino4 area of ±1, ±3, and $\pm5^oC$. The complete collection of both sets of MRF9 integrations available at twice-daily resolution is 570 members. Anomalies were derived by removing a least squares fit to the first three annual harmonics of the daily-averaged climatological-SST ensemble.

We also make use of two sets of AMIP-style AGCM integrations at monthly resolution, one from the MRF9 and the other from the NCAR CCM3.0 (Table 1). A 13-member ensemble of the MRF9 was integrated from 1950-1994 with the observed monthly SSTs specified globally (commonly referred to as GOGA). A 9-member ensemble was also integrated with the observed monthly tropical Pacific SSTs specified and climatological SSTs specified elsewhere (POGA). The second set of integrations comes from the NCAR CCM3.0. Kiehl et al. (1998) describe the CCM3.0 model in detail. A 12-member GOGA ensemble of CCM3.0 was integrated from 1950-1999. An 11-member ensemble of CCM3.0 was integrated from 1950-1999 with the observed monthly tropical (30^oN-30^oS) SSTs specified and climatological SSTs specified elsewhere (TOGA). Monthly anomalies were computed with respect to each model ensemble 1950-79 mean separately for the POGA, TOGA, and two GOGA integrations.

Analyzed height and vertical velocity fields were obtained from 50 years (1950-1999) of NCEP-NCAR reanalyses (Kistler et al. 1999) at twice-daily resolution. SST indices for Nino3 (5°N-5°S, 150°W-90°W) and Nino4 regions were constructed from the monthly Hadley Centre Sea Ice and Sea Surface Temperature (HadISST) dataset (Rayner et al. 2003). Anomalies of the reanalysis and HadISST data were computed by removing a least-squares fit to the first three annual harmonics of the 1950-79 data.

All AGCM and reanalysis anomaly fields were first smoothed to triangular truncation 31 using the spectral smoothing filter of Sardeshmukh and Hoskins (1984, hereafter SH). From the twice-daily data, the Fourier power spectrum of JFM 90-day anomaly segments was computed for each ensemble member (or each calendar year for the reanalyses) at every grid point. The 2 to 6.9 day band was summed over all 33 frequencies present to form the 2-7 day bandpass filtered variance. The variance fields were then spectrally smoothed to triangular truncation 12 using the SH filter to facilitate comparison with other studies using similar truncations (Whitaker and Sardeshmukh 1998, Chang and Fu 2002, 2003).

The smoothed 2-7 day 500 mb vertical velocity variance fields are referred to as the "omega stormtracks" and "stormtracks" where there is no possibility of confusion. Our spatial smoothing retains about 70% of the original standard deviation, but the pattern is preserved. The pattern correlation between the smoothed and unsmoothed omega stormtracks is 0.98.

Seasonal omega stormtrack anomaly fields were constructed by removing the 1950-79 JFM average from each JFM for the reanalysis data. The JFM mean of the NCEP MRF9 climatological SST ensemble mean was removed from each ensemble member of all 570 MRF9 experiments available at twice-daily resolution.

Empirical orthogonal functions (EOFs) of the 570 JFM stormtrack and 200 mb height anomaly fields were computed over the northern hemisphere (20-90°N), the western part of the northern hemisphere, the Pacific-North American (PNA) region (180-60°W, 20-90°N), and the North-Atlantic European region (60°W-60°E, 20-90°N) using the covariance matrix area-weighted by the cosine of latitude. From the EOFs, the equivalent spatial degrees of freedom (esdof) for all four domains were calculated using the method of Bretherton et al. (1999).

To orient the reader to the climatological pattern and interannual variance of the omega stormtracks compared to the 500 mb height stormtracks, the top panels of Fig. 3 show the standard deviation of 1950-79 2-7 day bandpass filtered 500 mb (left) omega and (right) height. The bottom panels show the interannual standard deviation of each stormtrack variable over 1950-99. The omega stormtracks shown in Fig. 3 (top left) are in excellent agreement with those

computed by Hoskins and Hodges (2002) using ECMWF data for 1979-2000. This agreement supports the findings of Compo et al. (2001) that the synoptic timescale vertical velocity variance is similar between the various observational estimates in the northern hemisphere extratropics. As illustrated in Fig. 3, the omega stormtracks are also particularly useful as they capture the well-known maxima in the Pacific and Atlantic and have a well-defined Mediterranean maximum seen in cyclone feature-tracking studies that is not found in climatologies of several other synoptically bandpass filtered variables such as 500 mb height (Hoskins and Hodges 2002).

As in Compo et al. (2001), we present anomalous stormtrack maps as variance differences rather than ratios to facilitate comparison with our previous work and to allow direct comparisons between predicted and observed stormtrack anomalies. Specifically, we present maps of

$$\Delta_\sigma = \mathrm{sgn}\left(\sigma_i^2 - \sigma_o^2\right) \times \left|\sigma_i^2 - \sigma_o^2\right|^{1/2}, \qquad (2)$$

where *i* indicates the year and *o* indicates climatological (or neutral) SST conditions. This quantity has the same units as the seasonal mean anomalies, is of comparable magnitude, and preserves the sign of the variance difference. The patterns of Δ_σ can also be directly interpreted and diagnosed in terms of the dynamical difference equations for second moment quantities.

3. Empirical Stormtrack model

a. Description of empirical storm track model

Understanding the connection between a background flow and the behavior of individual synoptic eddies evolving on it has long been a core problem in dynamical meteorology. The shift of focus to the link between a mean flow and the overall statistics of the synoptic eddies associated with it – "stormtracks" – is a relatively recent development (e.g., Blackmon et al. 1977; Lau 1988; Wallace et al. 1988; Farrell and Iannou 1994, 1995; Branstator 1995; Whitaker and Sardeshmukh 1998, Zhang and Held 1999). Whitaker and Sardeshmukh (1998) were able to simulate many aspects of the observed climatological winter-mean Pacific and Atlantic stormtracks given knowledge only of the climatological winter-mean flow at two levels in the upper and lower troposphere. Encouraged by this, they put their model to a harder test: to predict

the anomalous stormtracks for individual winters given the anomalous winter mean flow. Overall, Whitaker and Sardeshmukh had only limited skill at this, and the question remains whether this is due to nonlinearity of the mean-flow stormtrack relationship, the relative simplicity of the model, noise present in an individual winter mean flow and stormtracks, or some other factor.

Whitaker and Sardeshmukh's "dynamical stormtrack model" consists of stochastically perturbing a 2-level quasi-geostrophic model linearized about a specified mean flow to deduce the stormtracks associated with that flow. One can also think of constructing an "empirical stormtrack model", that uses a multiple linear regression operator estimated from data to predict the anomalous stormtracks associated with an anomalous mean flow. The prediction equation may formally be written as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \varepsilon \qquad (3)$$

where $\mathbf{G}$ is the linear regression operator, $\mathbf{x}$ (the predictor) is the anomalous mean flow, $\mathbf{y}$ (the predictand) is the anomalous stormtrack field, and ε is the error. For the northern hemisphere winter (January-March), we have constructed such a model in a truncated EOF space, with $\mathbf{x}$ as the anomalous mean 200 mb height and $\mathbf{y}$ as the anomalous omega stormtracks using the 570 NCEP AGCM integrations listed in Table 1.

The optimal $\mathbf{G}$ was determined by cross-validation, sequentially removing 30 members of the set of 570 at a time, computing $\mathbf{G}$ from the remaining 540, and then predicting the stormtrack anomalies in the excluded 30. All EOF truncations from 2 to 60 for 200 mb height and 2 to 70 for stormtrack anomalies were computed. The cross-validated root-mean-square error and average pattern correlation as a function of truncation are shown in Fig. 4. The cross-validated skill of the STM is not extremely sensitive to the choice of truncation. As such, we have chosen the truncation with the largest average pattern correlation: 40 EOFs of the 200 mb height field and 51 EOFs of the stormtrack anomaly field. In EOF space, $\mathbf{G}$ is then a 40 x 51 matrix.

b. Validation of STM

To establish the utility of a linear storm track model for diagnosing predictable stormtrack anomalies, one first has to demonstrate that it is accurate enough to predict a nonlinear GCM's ensemble-mean SST-forced stormtrack signal given the GCM's ensemble-mean seasonal flow. The ability of an empirical stormtrack model trained on an AGCM's *noise*, i.e. on unpredictable stormtrack and mean-flow variations, to predict SST-forced stormtrack variations cannot be inferred directly from Fig. 4, because (1) to the extent that the SST-forced mean-flow signal is weaker than the noise, one is putting the stormtrack model to an even harder test than in Fig. 4; and (2) the SST-forced mean-flow signal for an individual case may have features that are not captured in the truncated EOF space in which **G** operates. Fig. 5 shows that the STM is nevertheless successful at predicting the principal elements of the AGCM's ensemble-mean northern hemisphere stormtrack anomaly field in JFM 1987 given the AGCM's ensemble-mean seasonal mean flow anomaly. For this test, the EOFs and **G** itself were re-derived excluding the 60 members of the 1987 AGCM integrations. Over the hemisphere, the pattern correlation between the STM AGCM stormtrack anomalies and the fully nonlinear AGCM ensemble mean stormtrack anomalies is 0.9.

The diagnostic skill of the STM is also of interest. Is an empirical model trained on AGCM statistics relevant to the observations? As shown in the left hand panels of Fig. 5, this simple STM has considerable skill in diagnosing the pattern and amplitude of the observed anomalous winter-mean stormtracks of the 1987 El Nino winter in the NCEP Reanalysis dataset, given the observed anomalous winter-mean flow over the hemisphere. While there is an error of sign in the western Pacific, over the western part of the hemisphere, the pattern correlation is 0.7.

Fig. 6 shows similarly skillful results for the observed and AGCM predicted stormtrack anomalies of the 1989 La Nina. Here also, the EOFs and STM have been re-derived excluding all 60 members of the 1989 AGCM integrations. In the right hand panels, over the northern hemisphere, the pattern correlation between the STM-diagnosed AGCM stormtrack anomalies and the fully nonlinear AGCM ensemble-mean stormtrack anomalies is the same as for the 1987

case (0.9). The diagnostic skill seen in the left hand panels is also high, with pattern correlations over the hemisphere and its western part of 0.5 and 0.8, respectively.

Our STM is complementary to those recently developed by Chang and Fu (2003) and Peng et al. (2003). Our approach is to construct the STM from AGCM statistics to diagnose stormtrack anomalies from independent AGCM and observational mean-flow anomalies. Chang and Fu built a CCA between the anomalous mean flow and eddy statistics using the more recent part of the NCEP reanalysis dataset to assess the quality and decadal variability of eddy statistics in the earlier part. While successful in many respects, the CCA stormtrack model of Chang and Fu had a substantial problem with the amplitude of the diagnosed stormtrack anomalies. Peng et al. (2003) used multiple linear regression on an NCEP AGCM dataset to construct a linear operator relating monthly mean geopotential height to synoptic eddy vorticity fluxes for diagnosing the mean-flow/eddy feedback in that AGCM. Neither the stormtrack model of Peng et al. nor the STM used here (Figs. 5 and 6) have the amplitude problem of Chang and Fu, which probably arises from sample size limitations of the observed record.

For the remainder of the paper, the STM used is that derived from all 570 members of the twice-daily MRF9 integrations (Table 1).

4. Skill for predicting interannual stormtrack variations.

Having demonstrated that the STM can successfully diagnose both observed and AGCM ensemble-mean stormtrack anomalies in specific cases, we now use it to estimate the northern hemisphere stormtrack anomalies for the last 50 years from observations and 4 different sets of AGCM integrations that are independent of it. This will allow us to evaluate the diagnostic skill of the STM and to estimate the potential skill of predicting interannual stormtrack variations.

Fig. 7 shows, at each northern hemispheric gridpoint, the temporal anomaly correlation between the observed stormtrack anomalies and the STM-diagnosed stormtrack anomalies. The contours begin at the local 5% significance level of 0.25 and are plotted every 0.15 thereafter. In the top panel, the observed 200 mb height anomalies are given to the STM to diagnose the

stormtrack anomalies. In all other panels, AGCM ensemble mean 200 mb height anomalies are given to the STM to produce AGCM-predicted stormtrack anomalies. As suggested in Figs. 5 and 6, the top panel of Fig. 7 shows that the diagnostic skill of the STM, given the mean flow at only the 200mb level, is high over a large portion of the northern hemisphere, with a temporal correlation above 0.7 in several large areas. The failure of the STM over the western Pacific and south Asia, seen in Figs. 5 and 6, is also evident here. However, the STM accurately diagnoses most of the regions of largest interannual stormtrack variability (Fig. 3).

The bottom four panels of Fig. 7 show the actual skill in predicting interannual stormtrack anomalies. The stormtrack simulation skill has been estimated separately using the STM on ensemble mean 200 mb height anomalies from (left panels) GOGA integrations and (right panels) TOGA and POGA integrations using (middle panels) the NCAR CCM3.0 and (bottom panels) the NCEP MRF9 integrations described in Table 1. The correlations are very similar in the GOGA, TOGA, and POGA integrations, although a tendency for lower values in the latter two is apparent over the Gulf of Alaska and Europe. The results for the two AGCMs are generally consistent. The Pacific-only SST forcing contains almost all of the skill of the other integrations, providing evidence that most of the predictable stormtrack anomalies are forced by tropical Pacific SST anomalies.

To further characterize the interannual stormtrack variations, we examine the pattern correlation between observed and predicted stormtrack anomalies for each JFM season of 1950-99 in Fig. 8. As the diagnostic skill of the STM is high over certain parts of the hemisphere, the skill for simulating the pattern of stormtrack anomalies is calculated only for the Pacific-North American (PNA) and the North-Atlantic European (NATL-EUR) sectors. Fig. 8(c) shows the pattern correlations over the PNA sector between the observed stormtrack anomaly and that diagnosed from the STM using (black bars) the observed 200mb height anomaly, (green bars) GOGA CCM3.0, and (blue bars) TOGA CCM3.0. Results for the NCEP MRF9 are similar. The skill over the NATL-EUR region is shown in Fig. 8(d). The high correlations seen using the

observed 200mb height field show that the diagnostic skill of the STM is substantial in both regions.

Fig. 8 gives an idea of the case-dependence of stormtrack simulation skill. The GOGA and TOGA AGCM integrations both simulate the stormtrack anomalies with significant skill in many years over the PNA and NATL-EUR sectors. The case-to-case variations of skill are large, and not entirely attributable to sampling fluctuations. This skill is tested against the null hypothesis that the integrations are independent and have no relationship to the observed anomalies. In a monte carlo procedure using resampling with replacement, 5000 pattern correlations for the PNA sector and the NATL-EUR sector are computed between the observed stormtrack anomalies and random pairings of GOGA and TOGA predicted stormtrack anomalies. In a given monte carlo realization, both independently selected GOGA and TOGA maps are correlated with the same randomly-selected observed map. In only 5% of the monte carlo realizations did pairs of the two integrations simultaneously have pattern correlations exceeding ~0.2, shown by the thin horizontal line in Fig. 8(c) and (d). From sampling, over a 50-year period one would expect 3 pairs to exceed the significance threshold of ~0.2 simultaneously. In Fig. 8, 26 pairs of GOGA and TOGA integrations simultaneously exceed the significance test in the PNA sector, and 12 pairs exceed the threshold in the NATL-EUR sector. While 3 is the expected value, this is not the number of years that one would find in 5% or fewer of random sets of 50 years. That number should be higher.

Therefore, we have performed another, and much harder, test to determine the probability distribution for the number of pairs in a 50-yr set simultaneously exceeding the individual-year significance threshold of ~0.2. Using the same monte carlo procedure, but now with 5000 50-yr sets, we determine the number of years that both TOGA and GOGA correlations simultaneously exceed the threshold in each 50-yr set. From this, we find that over the PNA and NATL-EUR sectors, respectively, 11 and 9 years out of 50 randomly pass the threshold in 5% (250) of the random 50-yr sets. The CCM3.0 skills in Fig. 8 over the PNA and NATL-EUR regions are significant more often than expected from this test.

The timeseries in Fig. 8 suggest that tropical SST variations in the equatorial Pacific play a dominant role in the stormtrack skill over both the PNA and the NATL-EUR sectors. The AGCM has significant skill over the PNA sector in 26 winters. This result is inconsistent with the idea that predictable signals occur solely during moderate to strong ENSO events. Fig. 8(e) shows timeseries of winter-mean anomalies of SST in the Nino3 and Nino4 regions. PNA stormtrack anomalies are skillfully simulated in several years that are not usually classified as moderate to strong El Nino or La Nina events (e.g. NOAA Climate Prediction Center's subjective classification scheme at http://www.cpc.noaa.gov). For example, in JFM 1970 and 1991 both CCM3.0 runs have large skill, yet CPC considers both weak El Nino events.

Recently, Barsugli and Sardeshmukh (2002) presented evidence that the extratropical atmosphere may be more sensitive to SST anomalies in the Nino4 region than elsewhere in the tropical Indo-Pacific. In their study, the atmospheric response to a Nino4 anomaly is nearly twice that of the same magnitude Nino3 anomaly. For the stormtrack skill shown in Fig. 8, the dependence on Nino3 and Nino4 is quantified in Table 2. The pattern correlations are averaged based on the simultaneous magnitude of Nino3 and Nino4. Table 2 shows that the largest average stormtrack skill in both the PNA sector and the NATL-EUR sectors is obtained when both Nino3 and Nino4 have large amplitude. Interestingly, significant skill is also found in the PNA sector when Nino4 has large amplitude but Nino3 does not. These 9 cases represent an untapped set of potentially predictable responses to equatorial Pacific SSTA that may be captured by AGCMs but are hidden by stratifying data by Nino3 or Nino3.4 alone.

The results for the NATL-EUR sector in Table 2 are also interesting, pointing to some dependence on both Nino3 and Nino4 anomalies. The diagnostic skill of the STM is quite high throughout the record, consistent with the top panel of Fig. 7. The generally low simulation skill over the NATL-EUR sector supports the finding of other studies of AGCM skill predicting eddy variances in this region (e.g. Pavan and Doblas-Reyes 2000). The moderate skill when both Nino3 and Nino4 are large may be unexpected and suggests a sensitivity to the details of the SST forcing not revealed when stratifying skill by only one index of ENSO.

Fig. 8 and Table 2 suggest that the skill in forecasting stormtrack anomalies will depend on the details of the SST forcing. Such case dependence is not very useful unless the variations in skill can be anticipated and issued as part of the forecast. Fig. 9 suggests that they can.

Fig. 9 shows that the expected forecast skill for stormtrack anomalies can be significantly different from ENSO event to event. The expected skill ρ_{60} for (a) JFM 1987 and (b) JFM 1989 stormtrack anomalies is computed directly from Eq (1) using the AGCM predicted signal to noise ratio *S* from each 60 member ensemble. The expected skill can be significantly different between El Nino and La Nina. Note, for example, that the expected stormtrack prediction skill over northern Europe is greater than 0.55 for the 1987 case but less than 0.25 in the 1989 case. All differences of ~0.3 or more are significant above the 5% level, assuming *S* is distributed as a student-*t* statistic. It is also interesting to observe that the overall skill is expected to be higher during the La Nina case than the El Nino case over many regions not usually associated with an ENSO effect. The skill is greater for the 1989 case over the Atlantic, Africa, the Middle East, and South Asia. Examining only averages of skill will mask these substantial variations from case to case.

Comparing the expected skill in Fig. 9 for two particular forcing fields with the average actual skill in Fig. 7 from 50 different forcing fields further illustrates the case-dependence of skill. Given that Fig. 7 does not make the "perfect model" assumption, it is not surprising that its actual skill is generally lower than the expected skill in Fig. 9. What is perhaps more surprising is that in some regions (such as the Gulf of Mexico) the average actual skill is actually *higher* than the expected skill in Fig. 9. This again underscores the point that once one goes beyond the broad-brush similarities of Figs. 7 and 9, substantial regional differences become apparent. They highlight the case-to-case and GCM dependence of stormtrack predictability, with important implications for the local precipitation. We return to the important issue of whether the actual skill is related to the expected skill in section 6.

5. Skill for predicting decadal stormtrack variations.

Perhaps one can push an empirical linear stormtrack model harder still, and ask it to diagnose the decadal variability and 50-year stormtrack trend (discussed, e.g., in Hurrell and van Loon 1997; Chang and Fu 2002) given the observed JFM-mean flow. Towards this end, Chang and Fu (2003) had some success with their CCA stormtrack model derived from observed statistics, though the amplitude of the variations was weaker than observed. One would like to know whether the STM derived from AGCM seasonal anomalies is also relevant to the observed decadal fluctuations. Further, one would like to quantify how much of the decadal stormtrack variations are related to the SST forcing. These issues are addressed below.

a. 5-yr averages

The top panel of Fig. 10, similar to Fig. 7, shows the local temporal correlation between the observed 5-winter average stormtrack anomalies and those diagnosed by the STM given the 5-winter average 200 mb height anomaly. Note that the contouring now begins at the local 5% significance level of 0.4. The hemispheric coverage of significant diagnostic skill is much less than for the seasonal stormtrack. Most of the North Atlantic and Europe's decadal variations are accurately diagnosed with correlations over 0.85 in several regions. The STM also successfully diagnoses the 5-winter average stormtrack variations over the eastern Pacific, but not over large portions of western North America.

The bottom four panels of Fig. 10 suggest the degree of SST forcing of decadal stormtrack variations. They show the local temporal correlation between the observed 5-winter average stormtrack anomalies and those of the four AGCM integrations applying the STM to the 5-winter average ensemble-mean 200mb height anomalies. These skill maps are very similar to each other and largely consistent with the interannual skill in Fig. 7 but with decreased skill over the southern US and increased skill over northern Canada. A notable difference is the presence of significant skill in the southern US and western Atlantic in the CCM3.0 panels compared to the MRF9. Livezey et al. (1997) demonstrated that the MRF9 has substantial difficulty reproducing observed upper-level features in the North Atlantic, and this would adversely affect

our diagnosed stormtracks. The MRF9's poorer showing over the North Atlantic is also somewhat evident in Fig 7, but a known mass-leak in upper-level fields (Livezey et al. 1997) may further degrade performance for decadal variability in this region.

Time-varying pattern correlations between the simulated and observed 5-winter average stormtrack anomalies over the PNA and NATL-EUR sectors are shown in Fig. 11 using the CCM3.0 GOGA and TOGA integrations. Also shown are the standardized 5-winter-average JFM values for Nino3 and Nino4 SST anomalies. Over the PNA sector, the diagnostic skill of the STM is high for the last 20 years of the record and consistently low in the mid-1960s. For this region, the GOGA and TOGA CCM3.0 stormtracks have a similar variation in skill, suggesting that the 5-winter averages for much of the record are being forced by SSTs in the tropics. The high skill over the PNA sector appears to correspond with the periods of large magnitude in the 5-winter averages of Nino3 and Nino4, leading to the speculation that the decadal variability in ENSO itself is driving most of this skill variation. In contrast, over the NATL-EUR sector, the diagnostic skill is relatively high throughout the record, while the CCM3.0 integrations have low skill from the mid 1960s to mid 1980s.

Fig. 11 suggests that the decadal stormtrack anomalies over both the PNA and NATL-EUR sectors are SST-forced in several epochs. However, over the PNA sector, the skill is statistically significant for the record, while over the NATL-EUR sector it is not. The AGCM skill is compared to the null hypothesis that 5-winter averages of both runs are independent and have no relationship to the observed anomalies. A similar monte carlo procedure to that described for Fig. 8 is performed with 5000 resamplings with replacement of the observed and predicted 5-winter average storm tracks. From it, we find only 5% of both integrations exceed pattern correlations of ~0.3, shown by the thin horizontal line in Fig. 11(c) and (d). In a 46-yr sample of 5-winter averages, an additional 5000 resamplings showed that only 14 and 10 years would randomly pass the threshold more than 5% of the time in the PNA and NATL-EUR sectors, respectively. The GOGA and TOGA CCM3.0 skills in Fig.11 simultaneously exceed the threshold 27 times over the PNA sector but only 4 times over the NATL-EUR sector.

b. 50-yr trend

Our results for winter and 5-winter averages motivate us to use the AGCM data to examine the role of SST-forcing in even longer-term stormtrack variations. There has been some debate in the literature about the observed stormtrack changes over the last 50 years (e.g. Hurrell and van Loon 1997, Graham and Diaz 2001, Chang and Fu 2002, 2003, Harnik and Chang 2003). The long-term omega stormtrack variations and their relationship to previously studied stormtrack variables have not been addressed. Fig. 12 examines the 50-yr trend in the omega stormtracks and its relationship with the global SST forcing.

As the STM appears to have low diagnostic skill over most of the eastern part of the northern hemisphere (Fig. 10), our discussion of stormtrack trends focuses on the western half. The top panels of Fig. 12 show the linear 50-yr trend of JFM 200 mb height anomalies from (top left) NCEP reanalyses and (top right) CCM3.0 forced with global observed SSTs. Hoerling et al. (2001) discuss the pattern similarity of these two fields (pattern correlation of 0.86). The stormtrack trends diagnosed by the STM from these fields are shown in the middle panels. Their pattern correlation is 0.77 (significant above the 5% level assuming 8.3 esdof). The stormtrack trend calculated directly from the NCEP reanalyses is shown in the bottom panel. Both STM-estimated trends correlate with the observed trend higher than 0.6 (significant above the 10% level). This correspondence is less than found by Chang and Fu (2003) using CCA to diagnose the 300mb meridional wind stormtrack; in that study the predicted and observed trends correlated at 0.85 over the hemisphere.

Our STM-diagnosed reanalysis stormtrack trend in Fig. 12 is similar to several recent studies of long-term Pacific and Atlantic stormtrack variability in near-surface variables. Over the Pacific, the pattern in Fig. 12 is consistent with the trend pattern of 2 to 6 day sea level pressure variance analyzed by Graham and Diaz (2001), with decreasing variance over Alaska and western North America and increasing variance over the central north Pacific. Our STM-estimated trend pattern over western North America is also consistent with the decadal variability in 1000mb height found by Chang and Fu (2003).

Over the Atlantic sector, we find good agreement among the various trend estimates. The STM diagnosed stormtrack trends in Fig. 12 are very similar to those seen in the reanalysis stormtrack trends. They are also consistent with the 1000mb height stormtrack trends calculated by Gulev et al. (2002) and the 1000mb height decadal variability shown by Chang and Fu (2003).

Over western North America, there is substantial disagreement between the diagnosed and reanalysis storm trends in Fig. 12. Given the relatively good correspondence between lower tropospheric variance trends in previous studies and the STM-diagnosed omega stormtrack trends in this region, we concur with Harnik and Chang (2003) that the sign of the trend in upper-air stormtrack variables may be in error over western North America and the Pacific, with a possible cause the 1973 change in radiosonde reporting procedures (Kistler et al. 2001, Chang and Fu 2003). Another discontinuity is the introduction of satellite data into the NCEP reanalyses starting in 1979 (Kisler et al. 2001) whose effect on the analyzed synoptic variances has not been quantified.

In contrast, over the Atlantic-European sector, Harnik and Chang (2003) found that 300 mb meridional wind stormtrack trends derived from radiosondes were closer to those from the reanalysis in this region. Our results support the conclusion that the NCEP reanalysis upper-level stormtrack trends are more reliable over the Atlantic and eastern North America than over western North America or the Pacific.

The greater similarity of the CCM3.0's stormtrack trend with the STM-diagnosed rather than the reanalysis trend shown in Fig. 12 lends further evidence to support these conclusions. However, over the Atlantic, the CCM's stormtrack trend is shifted too far to the east, consistent with its eastward displacement of the observed high. The trend is also much weaker, consistent with the weaker amplitude of its height trend. The overall low amplitude of the height trend using global SST forcing, and the even lower amplitude using tropical SST forcing (see Hoerling et al. 2001 Fig. 3) suggests that less than half of the observed trend over the North Atlantic is consistent with the SST forcing, based on CCM3.0. This result may depend significantly on the AGCM used and needs to be investigated with other models.

6. Forecasting the forecast skill

The finding of significant variations of the expected skill in forecasting stormtrack anomalies is one of the central results of the present study. The case-to-case variations seen in Figs. 8 and 9 raise the possibility that there may be useful skill for an individual event different from the composite-based approach to estimating expected skill (e.g. Rowell 1998, Peng et al. 2000, Kumar et al. 2003). Such case-to-case variations in forecast skill are not helpful unless they also can be predicted (Anderson et al. 1999). Fig. 13 demonstrates that they can.

The thick curve in Fig. 13 shows the expected skill for a 12-member ensemble, assuming a perfect model, as a function of the signal to noise ratio S (Eq 1), reproduced from Fig. 2. One can also calculate the expected skill of a model with a time-varying systematic error (Sardeshmukh et al. 2000). As discussed in the Appendix, Sardeshmukh et al. determined that when a multivariate ensemble forecasting system has a systematic error, the expected skill is given by

$$\rho_n = S^2 / [(S^2+1)(S^2 + S_e^2 + n^{-1})]^{1/2}, \qquad (4)$$

where S_e is the ratio of the ensemble mean systematic error to the ensemble spread. The dotted curve shows the expected skill for a 12 member ensemble when a systematic error $S_e = 2S$ is present in the forecast (Eq 4). We have estimated separate S-values for every JFM of each of the four NCEP and NCAR AMIP-style integrations using the STM-diagnosed ensemble-mean stormtrack anomaly as the signal and combining 210 ensemble members from the 1987, 1989, and climatological MRF9 integrations (Table 1), with their respective ensemble means removed, to form the noise. The same spread is thus used for all S calculations. The symbols in Fig. 13 show the actual average simulation skill of the stormtrack anomalies. Averages are taken over similar S values for the PNA sector (filled circles) and the NATL-EUR sector (diamonds).

The stormtrack skill in Fig. 13 is consistent with the estimated S values in the Pacific sector but not in the Atlantic sector. A similar result is obtained for 5-yr averages but with larger error bars (not shown). These results are robust whether the tropically-forced runs are included or excluded (not shown). These results are also robust when the NCEP MRF9 and NCAR

CCM3.0 are considered separately (not shown). The systematic error over the Atlantic sector is on the order of twice the AGCMs' predicted signals. Because the stormtrack signals are obtained from the STM, the error must lie either in the STM or the specified 200 mb height anomaly. We believe the error is in the latter. The diagnostic skill of the STM is quite high over the Atlantic sector (Figs. 7 and 8) and is nearly the same on average over the PNA and NATL-EUR regions (Table 2). The STM also successfully recovers MRF9 ensemble-mean stormtrack anomalies over the Atlantic given only ensemble-mean 200 mb height anomalies in the 1987 (Fig. 5) and 1989 (Fig. 6) cases. These results demonstrate that both the MRF9 and CCM3.0 have a substantial systematic error over the North Atlantic-European sector in the 200mb height response to specified global and tropical SSTs.

Several previous studies have shown that many AGCMs have low skill predicting upper-level circulation and precipitation anomalies over the North Atlantic-European sector (e.g. Livezey et al. 1997, Brankovic and Palmer 2000, Doblas-Reyes et al. 2000, Graham et al. 2000, Peng et al. 2000, Shukla et al. 2000). The estimated *S* values in Fig. 13 and the expected skill shown in Fig. 9 suggest that, if the systematic error is correctable, the actual skill over the North-Atlantic Europe sector could be much higher than found here for stormtrack anomalies and, by inference, for other related quantities such as winter-mean heights and precipitation. Note that the systematic error diagnosed in Fig. 13 is not a constant that can be linearly removed, but is time-varying, making its correction much more difficult.

7. Conclusions.

Our study shows that there is indeed a predictable SST-forced stormtrack signal over much of the Northern Hemisphere in boreal winter, but one that may differ substantially from case to case and between El Nino and La Nina events. The signal is modest, but has large implications for the prediction of seasonal precipitation. The results further suggest that a predictable stormtrack signal may exist over the Pacific-North America sector in weak and even non-ENSO winters. The signal is sufficiently robust that it can be predicted *a priori*, given only

the SSTs in the tropics. Specifying extratropical SSTs in addition enhances the skill slightly. Decadal stormtrack variations are largely consistent with the tropical SST forcing, particularly in the second half of the record.

These conclusions do not apply in the North Atlantic-European sector. While a potentially predictable stormtrack signal may exist, a substantial systematic error has been diagnosed, on the order of twice the signal strength. If the error is correctable, the expected skill calculations of Figs. 9 and 13 suggest that the potential skill for this region may be much higher than the actual skill found here.

Our study has taken advantage of two relationships. First, the signal-to-noise ratio *S* and its related expected correlation skill ρ_n can be calculated for any multivariate quantity, for any forecasting situation. *S* then serves as a useful simple measure of predictability. When *S* is small, the skill of deterministic ensemble-mean predictions is low, and can easily be compromised by model errors and the use of small ensembles. With large ensembles and small model error, useful *probabilistic* predictions, especially of the altered risks of extreme anomalies, may still be possible, as stressed by Sardeshmukh et al. (2000).

Second, as shown by the success of recent empirical STMs (Chang and Fu 2003) and our own calculation, the relationship between the observed mean-flow and observed extratropical stormtrack anomalies is essentially linear and lies in a low-dimensional space. Our linear storm track model (STM) can reproduce a nonlinear GCM's storm track response to ENSO forcing, given only the GCM's 200 mb height response. These results indicate room for improvement in dynamical stormtrack models' simulation of the observed eddy variances, covariances, and fluxes.

The two relationships together have been used to estimate the local and regional predictability of winter-mean and 5-winter-mean storm track anomalies. The analysis of actual versus expected skill over the North-Atlantic European sector suggests that the poor skill seen in many studies of seasonal anomalies over the European sector may not represent an inherent predictability limit but rather a common systematic error that would be especially evident in

precipitation. Over the PNA sector, the results of Fig. 13 imply that the time variation of the actual skill seen in Fig. 8 may be predicted. An ensemble on the order of 128 members should be adequate to allow reliable estimation of *S* changes of 0.25 from winter to winter.

Acknowledgements.

The authors would like to acknowledge useful discussions with colleagues at CDC, particularly M. Alexander, J. Barsugli, M. Newman, C. Penland, and J. Whitaker. D. Hooper, C. Smith, and G. Bates provided invaluable assistance with data processing. The HadISST data were kindly provided by N. Rayner and the Hadley Centre of the Met Office. NCEP-NCAR reanalysis data were obtained from the CDC archives (http://www.cdc.noaa.gov). NCEP MRF9 AMIP runs were made courtesy of M. Ji of NCEP. NCAR CCM3.0 runs were provided courtesy of Y.-H. Lee of NCAR. This work was partially supported by a grant from NOAA's Clivar-Pacific Program.

Appendix: Derivation of multivariate predictability as a function of the signal to noise ratio.

One way to increase the value of climate predictions is to issue the expected skill of the forecast as part of the prediction. van den Dool and Toth (1991) derived the expected value of a forecast's correlation skill when predicting the ensemble mean anomaly of any univariate distribution using an infinite member ensemble. Rowell (1998) extended van den Dool and Toth's result to the use of a *n*-member ensemble forecast. Kumar and Hoerling (2000) derived the results of Rowell (1998) for the specific case of forecasting the sign of the mean anomaly of a Gaussian distribution.

Sardeshmukh et al. (2000) further developed the results of Rowell (1998) for a multivariate forecast that has errors. They derived the expected skill for an ensemble forecasting system predicting any multivariate quantity that has defined first and second moments. The forecasting system need not have a perfect model. Following Sardeshmukh et al., consider a

multivariate distribution $P_m(\langle \mathbf{x}+\mathbf{x_e}\rangle, \mathbf{C_{0m}})$ that represents the altered model probability density function (PDF) of some quantity and $P(\langle \mathbf{x}\rangle, \mathbf{C_0})$ that is the true PDF of that quantity, such as winter stormtrack anomalies during an El Nino event. Here $\langle \mathbf{x}\rangle$ is the population mean anomaly state vector, $\langle \mathbf{x}+\mathbf{x_e}\rangle$ is model's population mean anomaly state vector, where $\langle \mathbf{x_e}\rangle$ equals the model's error in predicting the population mean, $\mathbf{C_0}$ is the covariance matrix of the variations $\mathbf{x}'$ around $\langle \mathbf{x}\rangle$, and $\mathbf{C_{0m}}$ is the model's covariance matrix of variations $\mathbf{y}'$ about $\langle \mathbf{x}+\mathbf{x_e}\rangle$. Note that P and P_m can be any multivariate distributions that have defined first and second moments. Also, note that none of the parameters of these distributions need necessarily be the same for El Nino and La Nina or even from case-to-case. The PDF of ensemble mean forecasts issued from an *n*-member ensemble with this model is $P_m(\langle \mathbf{x}+\mathbf{x_e}\rangle, n^{-1}\,\mathbf{C_{0m}})$. Assume that a vector $\mathbf{y}$ is issued as the forecast, and the real atmosphere picks a vector $\mathbf{x} = \langle \mathbf{x}\rangle + \mathbf{x}'$ from P as its stormtrack anomaly field. The average anomaly correlation of the observed and predicted vectors is then

$$\begin{aligned}\rho_n &= \langle \mathbf{x}\cdot\mathbf{y}\rangle \Big/ \left(\langle \mathbf{x}\cdot\mathbf{x}\rangle\langle \mathbf{y}\cdot\mathbf{y}\rangle\right)^{1/2} \\ &= \langle \mathbf{x}\rangle\cdot\langle \mathbf{x}\rangle \\ &\div\left[\left(\langle \mathbf{x}\rangle\cdot\langle \mathbf{x}\rangle + \langle \mathbf{x}'\cdot\mathbf{x}'\rangle\right)\left(\langle \mathbf{x}\rangle\cdot\langle \mathbf{x}\rangle + \langle \mathbf{x}_e\rangle\cdot\langle \mathbf{x}_e\rangle + n^{-1}\langle \mathbf{y}'\cdot\mathbf{y}'\rangle\right)\right]^{1/2}\end{aligned} \tag{A1}$$

where we have assumed that $\langle \mathbf{x}\rangle\cdot\langle \mathbf{x}_e\rangle = 0$. The dot product here represents a general scalar product of the form $\mathbf{x}\cdot\mathbf{y} = \mathbf{x}^T\mathbf{W}\mathbf{y}$, where $\mathbf{W}$ is any suitable positive-definite weight matrix. Note that $\langle \mathbf{x}'\cdot\mathbf{x}'\rangle = \mathrm{Tr}\left[\mathbf{W}^{1/2}\mathbf{C}_0\mathbf{W}^{1/2}\right]$ and $\langle \mathbf{y}'\cdot\mathbf{y}'\rangle = \mathrm{Tr}\left[\mathbf{W}^{1/2}\mathbf{C}_{0m}\mathbf{W}^{1/2}\right]$. The weight matrix $\mathbf{W}$ can be chosen to emphasize a particular gridpoint (as in Fig. 9), a linear combination of variables over a region (as in Fig. 13), or be set equal to identity to examine skill over the entire atmosphere.

If we further assume that the model correctly reproduces the second moment, i.e., $\mathbf{C_{0m}} = \mathbf{C_0}$, then (A1) becomes (4):

$$\rho_n = S^2 \Big/ \left[\left(S^2+1\right)\left(S^2 + S_e^2 + n^{-1}\right)\right]^{1/2}, \tag{A2}$$

where $S = \left[\langle \mathbf{x}\rangle\cdot\langle \mathbf{x}\rangle \big/ \langle \mathbf{x}'\cdot\mathbf{x}'\rangle\right]^{1/2}$ and $S_e = \left[\langle \mathbf{x}_e\rangle\cdot\langle \mathbf{x}_e\rangle \big/ \langle \mathbf{x}'\cdot\mathbf{x}'\rangle\right]^{1/2}$. For a "perfect" model, both $\langle \mathbf{x_e}\rangle = 0$ and $\mathbf{C_{0m}} = \mathbf{C_0}$, and (A1) leads to (1):

$$\rho_n = S^2 \Big/ \left[\left(S^2+1\right)\left(S^2 + n^{-1}\right)\right]^{1/2}. \tag{A3}$$

In the limit as the ensemble size goes to infinity, for a perfect model $\rho_\infty = S \Big/ \sqrt{\left(S^2+1\right)}$. For univariate distributions, ρ_∞^2 is closely related to the predictability measure examined by Koster et al. (2000).

Graham et al. (2000) found that increasing number of ensemble members did little to increase skill, even in the context of a perfect model. This can be understood directly in terms of ρ_n in (A3) and graphically as illustrated in Fig. 2. Graham et al. (2000) empirically calculated ρ_n with n=9 and n= 18 and compared it to their actual anomaly correlation (their Figs. 13 and 14). It is clear from A3 and Fig. 2 that little is expected to be gained by increasing ensemble size from 9 to 18 members. The advantage of using much larger ($n \geq 128$) ensemble sizes lies in improving the ability to forecast the forecast skill. Note also from A2, that any advantage for actual skill can be lost from model error, as illustrated by the dotted curve in Fig. 13.

References

Anderson, J., H. van den Dool, A. Barnston, W. Chen, W. Stern, and J. Ploshay, 1999: Present-day capabilities of numerical and statistical models for atmospheric extratropical seasonal simulation and prediction. *Bull. Am. Met. Soc.*, **80**,1349-1361.

Barsugli, J.J, and P.D. Sardeshmukh, 2002: Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin. *J. Climate*, **15**, 3427-3442.

Blackmon, M.L., J.M. Wallace, N.-C. Lau, and S.L. Mullen, 1977: An observational study of the northern hemisphere wintertime circulation. *J. Atmos. Sci.*, **34**, 1040-1053.

Brankovic, C. and T.N. Palmer, 1997: Atmospheric seasonal predictability and estimates of ensemble size. *Mon. Wea. Rev.*, **125**, 859-874.

_____ and _____, 2000: Seasonal skill and predictability of ECMWF PROVOST ensembles. *Q. J. R. Meteorol. Soc.,* **126**, 2035-2067.

Branstator, G., 1995: Organization of stormtrack anomalies by recurring low-frequency circulation anomalies. *J. Atmos. Sci.*, **52**, 207-226.

Bretherton, C.S., M. Widmann, V.P. Dymnikov, J.M. Wallace, and I. Blade, 1999: The effective number of spatial degrees of freedom of a time-varying field. *J. Climate*, **12**, 1990-2009.

Carillo, A., P.M. Ruti, and A. Navarra, 2000: Storm tracks and zonal mean flow variability: a comparison between observed and simulated data. *Cli. Dyn.,* **16**, 219-228.

Chang, E.K.M., and Y. Fu, 2002: Interdecadal variations in Northern hemisphere winter storm track intensity. *J. Climate*, **15**, 642-658.

_____, and _____, 2003: Using mean flow change as a proxy to infer interdecadal storm track variability. *J. Climate*, **16***,* 2178-2196.

Chen, W.Y., and H.M. van den Dool, 1997: Atmospheric predictability of seasonal, annual, and decadal climate means and the role of the ENSO cycle: A model study. *J. Climate*, **10**, 1236-1254.

Compo, G.P., P.D. Sardeshmukh, and C. Penland, 2001: Changes of subseasonal variability associated with El Nino. *J. Climate*, **14**, 3356-3374.

Doblas-Reyes, F.J., M. Deque, and J.-P. Piedelievre, 2000: Multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.,* **126**, 2069-2087.

Farrell, B.F., and P.J. Iannou, 1994: A theory for the statistical equilibrium energy and heat flux produced by transient baroclinic waves. *J. Atmos. Sci*., **51**, 2685-2698.

_____, and _____, 1995: Stochastic dynamics of the midlatitude atmospheric jet. *J. Atmos. Sci*., **52**, 1642-1656.

Fraedrich, K., 1990: European Grosswetter during the warm and cold extremes of the El Nino/Southern Oscillation. *Int. J. Climatol*., **10**, 21-31.

_____, 1994: An ENSO impact on Europe - A review. *Tellus*., **46A**, 541-552.

_____, and K. Mueller, 1992: Climate anomalies in Europe associated with ENSO extremes. *Int. J. Climatology*, **12**, 25-31.

Graham, R.J., A.D.L. Evans, K.R. Mylne, M.S.J. Harrison, and K.B. Robertson, 2000: An assessment of seasonal predictability using atmospheric general circulation models. *Q. J. R. Meteorol. Soc.,* **126**, 2211-2240.

Graham, N.E., and H.F. Diaz, 2001: Evidence for intensification of north Pacific winter cyclones since 1948. *Bull. Am. Met. Soc*., **82**, 1869-1893.

Gulev, S.K., T. Jung, and E. Ruprecht, 2002: Climatology and interannual variability in the intensity of synoptic-scale processes in the North Atlantic from the NCEP-NCAR reanalysis data. *J. Climate*, **15**, 809-828.

Harnik, N. and E.K.M. Chang, 2003: Storm track variations as seen in radiosonde observations and reanalysis data. *J. Climate*, **16**, 480-495.

Hoerling, M.P., J.W. Hurrell, and T. Xu, 2001: Tropical origins for recent North Atlantic climate change. *Science*, **292**, 90-92.

_____, and M. Ting, 1994: Organization of extratropical transients during El Niño. *J. Climate*, **7**, 745–766.

Hoskins, B. J. and K.I. Hodges, 2002: New perspectives on the Northern Hemisphere winter storm tracks. *J. Atmos. Sci.*, **59**, 1041-1061.

Hurrell, J.W., and H. van Loon, 1997: Decadal variations of climate associated with the North Atlantic Oscillation. *Climatic Change*, **36**, 301-326.

Kanamitsu, M., and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6**, 425-435.

Kiehl, J.T., J.J. Hack, G. B. Bonan, B.A. Boville, D.L. Williamson, and P.J. Rasch, 1998: The National Center for Atmospheric Research Community Climate Model: CCM3.0. *J. Climate*, **11**, 1131-1150.

Kistler, R., and Coauthors, 2001: The NCEP-NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bull. Am. Met. Soc.,* **82**, 247-268.

Koster, R.D., M.J. Suarez, and M. Heiser: 2000: Variance and predictability of precipitation at seasonal-to-interannual timescales. *J. Hydrometeorology*, **1**, 26-46.

Kumar, A. and M.P. Hoerling, 1998: Specification of regional sea surface temperatures in atmospheric general circulation model simulations. *J. Geophys. Res*., *103*, 8901-8907.

_____, and _____, 2000: Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Am. Met. Soc.*, **81**, 255-264.

_____, _____, M. Ji, A. Leetma, and P. Sardeshmukh, 1996: Assessing a GCM's suitability for making seasonal predictions. *J. Climate*, **9**, 115-129.

_____, S.D. Schubert, and M.S. Suarez, 2003: Variability and predictability of 200-mb seasonal mean heights during summer and winter. *J. Geophys. Res*., **108 D5**, 4169-4178.

Lau, N.C., 1988: Variability of the observed midlatitude storm tracks in relation to low-frequency changes in the circulation pattern. *J. Atmos. Sci*., **45**, 2718-2743.

Livezey, R. E., M. Masutani, A. Leetmaa, H. Rui, M. Ji, and A. Kumar, 1997: Teleconnective response of the Pacific-North American region atmosphere to large central equatorial Pacific SST anomalies. *J. Climate*, **10**, 1787-1820.

Matthews, A. J., and G. N. Kiladis, 1999: Interaction between ENSO, transient circulation, and tropical convection over the Pacific. *J. Climate.,* **12**, 3062–3086.

May, W., and L. Bengtsson, 1998: The signature of ENSO in the northern hemisphere midlatitude seasonal mean flow and high-frequency intraseasonal variability. *Meteor. Atmos. Phys*., **69**, 81-100.

Pavan, V. and F.J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. *Cli. Dyn*., **16**, 611-625.

Peng, P., A. Kumar, A.G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and the Scripps-MPI ECHAM3 models. *J. Climate*, **13**, 3657-3679.

_____,_____, and H. van den Dool, 2002: An analysis of multimodel ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, **107**, 4710-4722.

Peng, S., W.A. Robinson, and S. Li, 2003: Mechanisms for the NAO response to the north Atlantic SST tripole. *J. Climate,***16**, 1987-2004.

Rayner, N.A., D.E. Parker, E.B. Horton, C.K. Folland, L.V. Alexander, D.P. Rowell, E.C Kent, and A . Kaplan, 2003: Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century. *J. Geophys. Res*., in press.

Rowell, D.P., 1998: Assessing potential seasonal predictability with an ensemble of multi-decadal GCM simulations. *J. Climate*, **11**, 109-120.

Sardeshmukh, P.D., and B.J. Hoskins, 1984: Spatial smoothing on the sphere. *Mon. Wea. Rev.*, **112**, 2524-2529.

_____, G. P. Compo, and C. Penland, 2000: Changes of probability associated with El Nino. *J. Climate*, **13**, 4268-4286.

Schubert, S.D., M.J. Suarez, Y. Chang, and G. Brantstator, 2001: The impact of ENSO on extratropical low-frequency noise in seasonal forecasts. *J. Climate*, **14**, 2351-2365.

Shukla, J., and Coauthors, 2000: Dynamical seasonal prediction. *Bull. Am. Met. Soc.*, **81**, 2593-2606.

Smith, C.A., and P.D. Sardeshmukh, 2000: The effect of ENSO on the intraseasonal variance of surface temperatures in winter. *Int. J. Climatol.*, **20**, 1543-1557.

Straus, D.M., and J. Shukla, 1997: Variations of midlatitude transient dynamics associated with ENSO. *J. Atmos. Sci.*, **54**, 777-790.

van den Dool, H.M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. and Forecasting*, **6**, 76-85.

Wallace, J.M., G.-H. Lim, and M. Blackmon, 1988: Relationship between cyclone tracks, anticyclone tracks, and baroclinic waveguides. *J. Atmos. Sci*., **45**, 439-462.

The WASA Group, 1998: Changing waves and storms in the northeast Atlantic? *Bull. Amer. Meteor. Soc.*, **79**, 741-760.

Whitaker, J.S., and P. D. Sardeshmukh, 1998: A linear theory of extratropical synoptic eddy statistics. *J. Atmos. Sci,* **55**, 237-258.

Zhang, Y. and I.M. Held, 1999: A linear stochastic model of a GCM's midlatitude storm tracks. *J. Atmos. Sci,* **56**, 3416-3436.

Zwiers, F.W., X.L. Wang, and J. Sheng, 2000: Effects of specifying bottom boundary conditions in an ensemble of atmospheric GCM simulations. *J. Geophys. Res.*, **105**, 7295-7315.

Table captions

Table 1: Integrations of Atmospheric General Circulation Models available at twice-daily and monthly resolution used in the present study

Table 2: Average storm track pattern correlation between the observed winter-mean storm track and that predicted by the STM given 200mb height anomaly fields from NCEP-NCAR reanalysis (OBS) and ensemble-mean anomalies of CCM3.0 tropical SST forced (TOGA) and global SST forced (GOGA) integrations from 1950 to 1999. The skill is stratified by the magnitude of NINO3 and NINO4 indices and averaged separately over the Pacific-North America (PNA) and North Atlantic-European (NATL-EUR) regions. Average correlations significant at or above the 5% level are indicated by bold italics.

Figure captions

Figure 1**.** Signal to noise ratio *S* from the JFM 1987 El Nino for (a) seasonal mean 500 mb vertical velocity, (b) seasonal 2-7 day bandpass variance of vertical velocity, and (c) seasonal mean precipitation. The contour interval is 0.2. The zero contour has been suppressed. The 10% significance level is 0.22 using a 2-sided *t* test. All plots are field significance at the 5% level assuming at least 15 esdof. Pattern correlations between the respective fields are indicated next to the arrows.

Figure 2. Expected anomaly correlation skill $\rho_{\boldsymbol{n}}$ of forecasts made from the mean of $\boldsymbol{n}$=1, 12, 25, 60, and infinite member ensembles as a function of the signal-to-noise ratio $\boldsymbol{S}$. Thickened portions of curves illustrate uncertainty in expected skill $\rho_{\boldsymbol{n}}$ for $\boldsymbol{S}$ =0.7 due to uncertainty from estimating $\boldsymbol{S}$ using an $\boldsymbol{n}$-member ensemble, assuming that $\boldsymbol{S}$ is distributed as a student-t statistic. [Adapted from Sardeshmukh et al. 2000].

Figure 3. (top) Seasonal mean and (bottom) interannual standard deviation of 2 to 7 day bandpass variance of (left) 500 mb vertical velocity and (right) 500 mb height variance from NCEP-NCAR reanalyses. The square root of each field is plotted. Contour intervals are (top left) 0.01 Pa s^{-1} (top right) 5 m, (bottom left) 0.005 Pa s^{-1}, (bottom right) 2.5 m.

Figure 4. Cross-validated skill of the empirical storm track model as function of the number of retained EOFs of January-March 200 mb height seasonal mean and 500 mb vertical velocity bandpass variance seasonal anomalies. (a) Normalized error. (b) Anomaly pattern correlation. Contour interval is 0.05 in both panels with the addition of (a) 0.623, 0.625 and (b) 0.595, 0.598 contours. Shading begins at (a) 0.625, (b) 0.595. Horizontal and vertical lines show the truncation used for the storm track model described in the text.

Figure 5: Seasonal anomalies for January-March 1987 from (left) NCEP reanalyses and (right) AGCM ensemble of 60 members forced with 1987 observed SSTs. (top) 200 mb height anomaly. (middle) 500 mb storm track (vertical velocity 2-7 day band passed variance anomaly) diagnosed using the empirical linear storm track model. (bottom) 500 mb storm track. The plotted quantity in the middle and bottom panels is the signed square root of the variance anomaly. Contour intervals are (top) 20m and (middle and bottom) 0.01 Pa/s with the zero contour suppressed. (top) Light shading indicates negative anomalies. (middle and bottom) Dark (light) shading indicates positive (negative) anomalies.

Figure 6. Same as Fig. 5 but for 1989 January-March seasonal anomalies.

Figure 7. Predictability of storm tracks estimated using the storm track model. Green shading begins at the 5% significance level of 0.25. Contour interval is 0.15 thereafter.

Figure 8. Timeseries of anomaly pattern correlation for January-March (JFM) storm track anomalies. (a) and (b) show regions used in subsequent panels. (c) and (d) Pattern correlation over (c) the Pacific-North American (20-90N, 180-60W) and (d) North Atlantic-European sectors (20-90N, 60W-60E) of observed and STM diagnosed storm track anomalies using 200 mb height JFM anomalies from (black bars) observed, (green bars) CCM3 AGCM forced with global SSTs, (blue bars) CCM3 AGCM forced with tropical SSTs. Thin horizontal line shows the 5% significance levels for both AGCM integrations having skill. (e) Timeseries of JFM Nino3 (orange and light blue) and Nino4 (red and dark blue) normalized by their respective standard deviations.

Figure 9. Case dependence of the predictability of storm tracks. Comparison of expected local anomaly correlation skill for (a) 1987 and (b) 1989 global SST forcing estimated using 60-

member ensembles to calculate the signal-to-noise ratio S and then applying Eq (1) to calculate ρ_{60}. Contour interval is 0.15 starting at 0.25.

Figure 10. Skill for 5-winter averages. Green shading begins at the 5% significance level of 0.4. Contour interval is 0.15 thereafter.

Figure 11. As in Fig. 9 but for 5-winter averages.

Figure 12. Linear trends for 1950-99. Plotted as the change over 50-years. (top panels) Trend in 200 mb heights from (left) NCEP-NCAR reanalysis and (right) CCM3 GOGA. Contour interval is 20m. Negative trends are shaded. (middle and bottom) Trend in the omega storm track from the STM using 200 mb heights from (left) NCEP-NCAR reanalyses and (right) CCM3 GOGA. (bottom) Trend in the stormtrack from NCEP-NCAR reanalyses. Contour intervals is 0.01 Pa s^{-1}, with the zero contour suppressed. Light (dark) regions indicate negative (positive) trends.

Figure 13. Anomaly correlation skill of stormtrack forecasts made using the CCM3 and MRF9 diagnosed stormtracks for January-March season. Solid curve shows the expected correlation skill ρ_n of forecasts made from the mean of n=12 member ensembles as a function of the signal to noise ratio S based on Eq (1). Dotted curve shows the expected skill ρ_{12} when a systematic error S_e=2S is present in the forecast based on Eq (4). Symbols show the actual skill of storm track forecasts for the PNA region (filled circles) and North Atlantic-European region (diamonds) binned over similar S values. Bins widths are 0.25 from S=0 to S=1 and 0.5 thereafter. Percentage of cases in each bin is indicated. Error bars show the 95% confidence interval using the Fisher z-transformation and assuming 6 esdof.

Table 1: Integrations of Atmospheric General Circulation Models available at twice-daily and monthly resolution used in the present study

NCEP MRF 9 T40L18 (twice-daily)	Global SSTs	Climatological 1987 Jan-March 1989 Jan-March	90 members 60 members 60 members	Sardeshmukh et al. 2000 Compo et al. 2001
	NINO4 anomaly	±1 Jan-March ±3, ±5 Jan-March	90 members each 45 members each	This study
NCEP MRF 9 T40L18 (AMIP-style, monthly)	Global SSTs (GOGA)	1951-1994	13 members	Livezey et al. 1997
	Pacific SST anomaly (POGA)	1951-1994	9 members	Kumar and Hoerling 1998
CCM3.0 T42L18 (AMIP-style, monthly)	Global SSTs (GOGA) Tropical SST anomaly (TOGA)	1950-1999 1950-1999	12 members 11 members	Kiehl et al. 1998

Table 2: Average storm track pattern correlation between the observed winter-mean storm track and that predicted by the STM given 200mb height anomaly fields from NCEP-NCAR reanalysis (OBS) and ensemble-mean anomalies of CCM3.0 tropical SST forced (TOGA) and global SST forced (GOGA) integrations from 1950 to 1999. The skill is stratified by the magnitude of NINO3 and NINO4 indices and averaged separately over the Pacific-North America (PNA) and North Atlantic-European (NATL-EUR) regions. Average correlations significant at or above the 5% level are indicated by bold italics.

	$\|NINO3\| \geq 1\sigma$	$\|NINO3\| < 1\sigma$	$\|NINO3\| \geq 1\sigma$	$\|NINO3\| < 1\sigma$
	$\|NINO4\| \geq 1\sigma$	$\|NINO4\| \geq 1\sigma$	$\|NINO4\| < 1\sigma$	$\|NINO4\| < 1\sigma$
Number of Cases	11	9	2	28
PNA				
TOGA	***0.68***	***0.39***	0.13	0.11
GOGA	***0.61***	***0.36***	0.23	***0.25***
OBS	***0.72***	***0.57***	***0.65***	***0.50***
NATL-EUR				
TOGA	***0.29***	0.11	-0.12	0.05
GOGA	***0.30***	0.19	-0.24	0.13
OBS	***0.57***	***0.57***	0.27	***0.57***

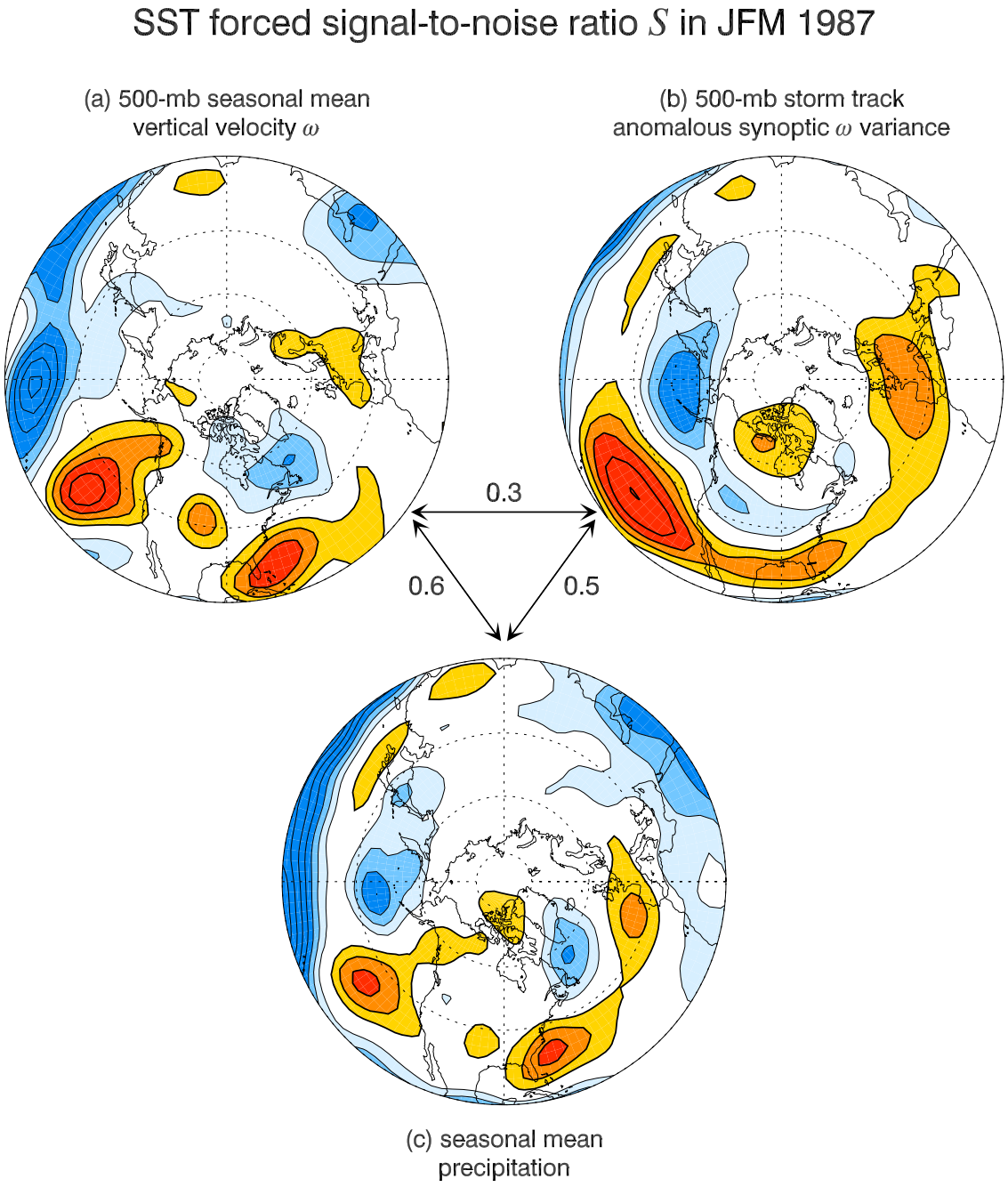

Figure 1. Signal to noise ratio S from the JFM 1987 El Nino for (a) seasonal mean 500 mb vertical velocity, (b) seasonal 2-7 day bandpass variance of vertical velocity, and (c) seasonal mean precipitation. The contour interval is 0.2. The zero contour has been suppressed. The 10% significance level is 0.22 using a 2-sided t test. All plots are field significance at the 5% level assuming at least 15 esdof. Pattern correlations between the respective fields are indicated next to the arrows.

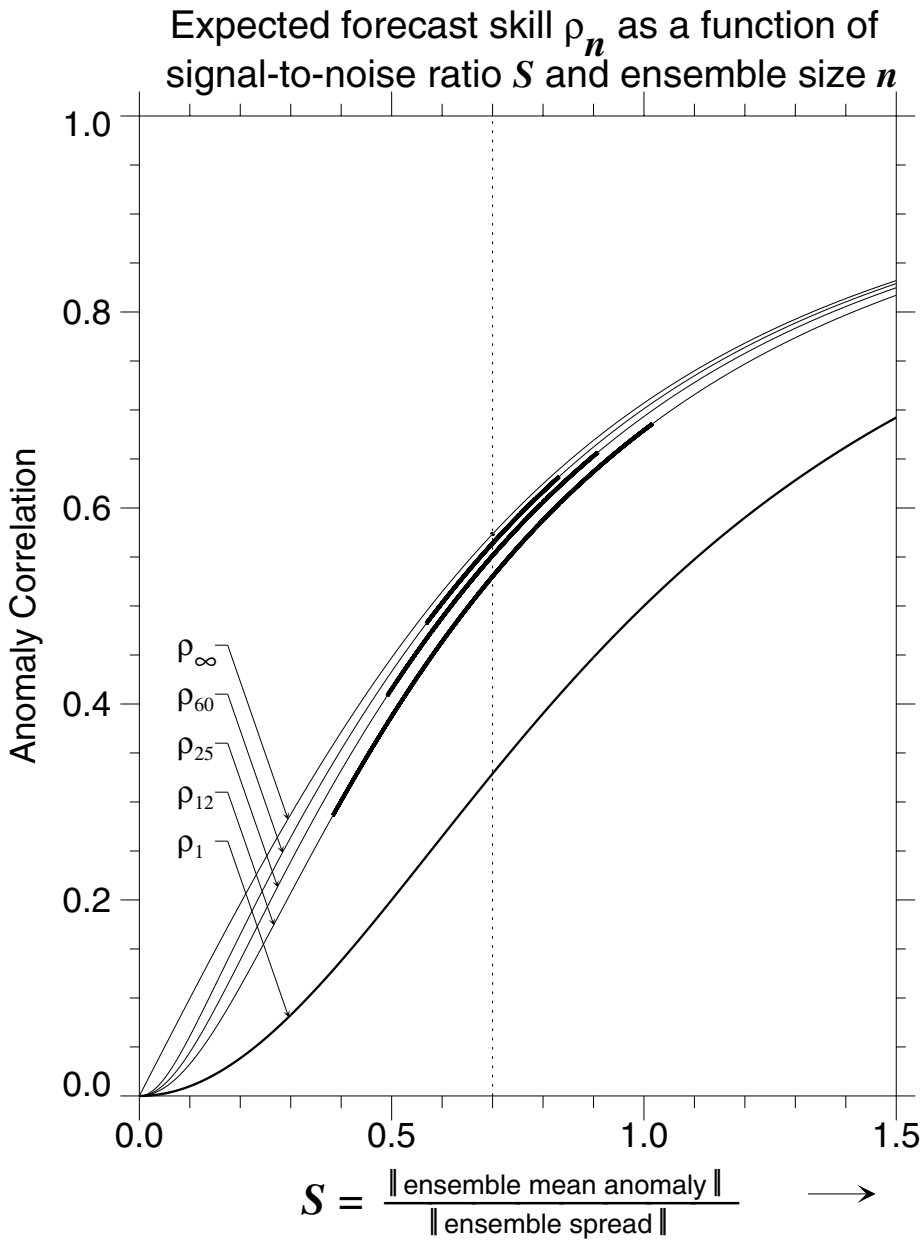

Figure 2. Expected anomaly correlation skill ρ_n of forecasts made from the mean of n=1, 12, 25, 60, and infinite member ensembles as a function of the signal-to-noise ratio S. Thickened portions of curves illustrate uncertainty in expected skill ρ_n for S =0.7 due to uncertainty from estimating S using an n-member ensemble, assuming that S is distributed as a student-t statistic. [Adapted from Sardeshmukh et al. 2000].

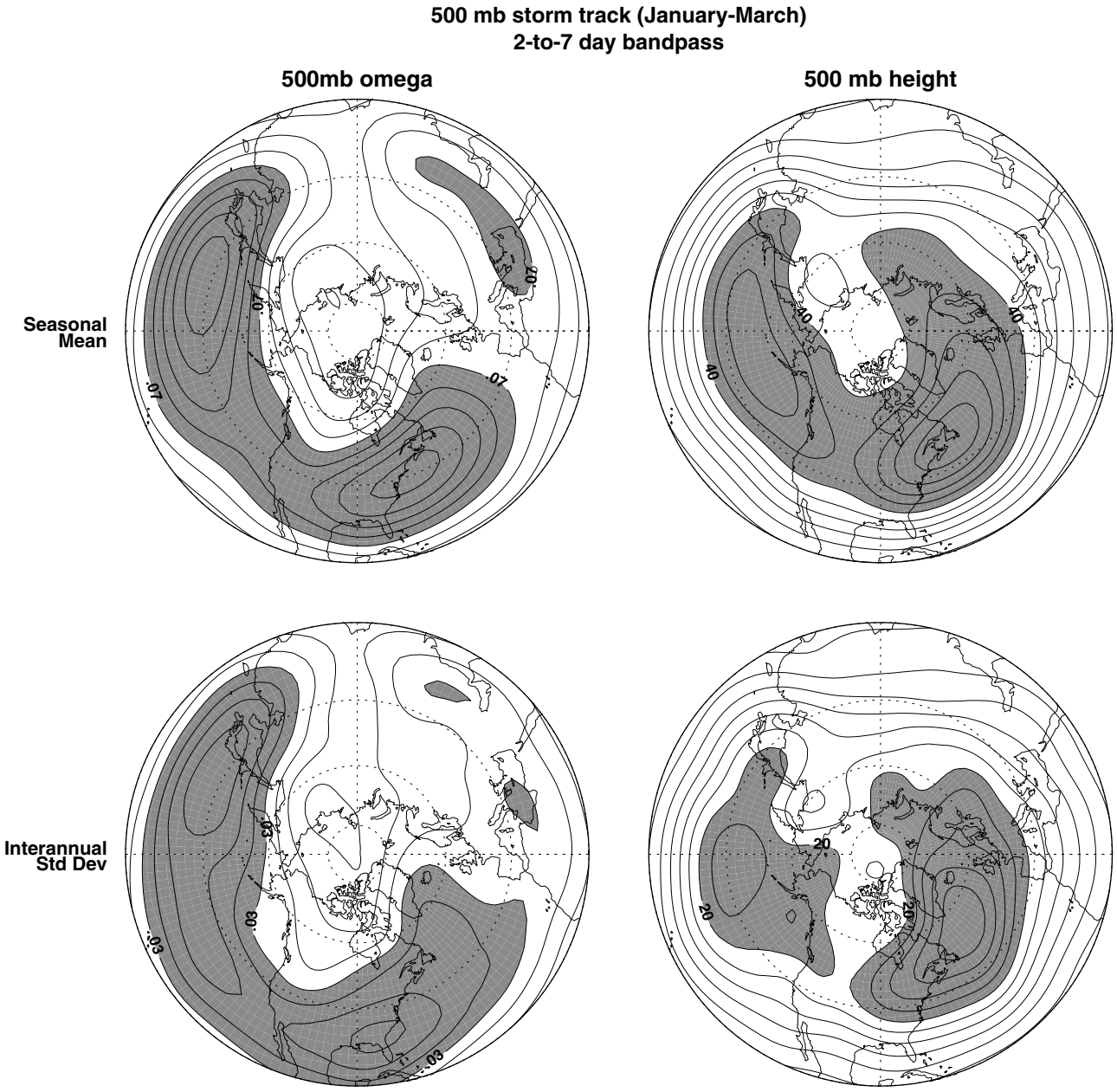

Figure 3. (top) Seasonal mean and (bottom) interannual standard deviation of 2 to 7 day bandpass variance of (left) 500 mb vertical velocity and (right) 500 mb height variance from NCEP-NCAR reanalyses. The square root of each field is plotted. Contour intervals are (top left) 0.01 Pa s^{-1} (top right) 5 m, (bottom left) 0.005 Pa s^{-1}, (bottom right) 2.5 m.

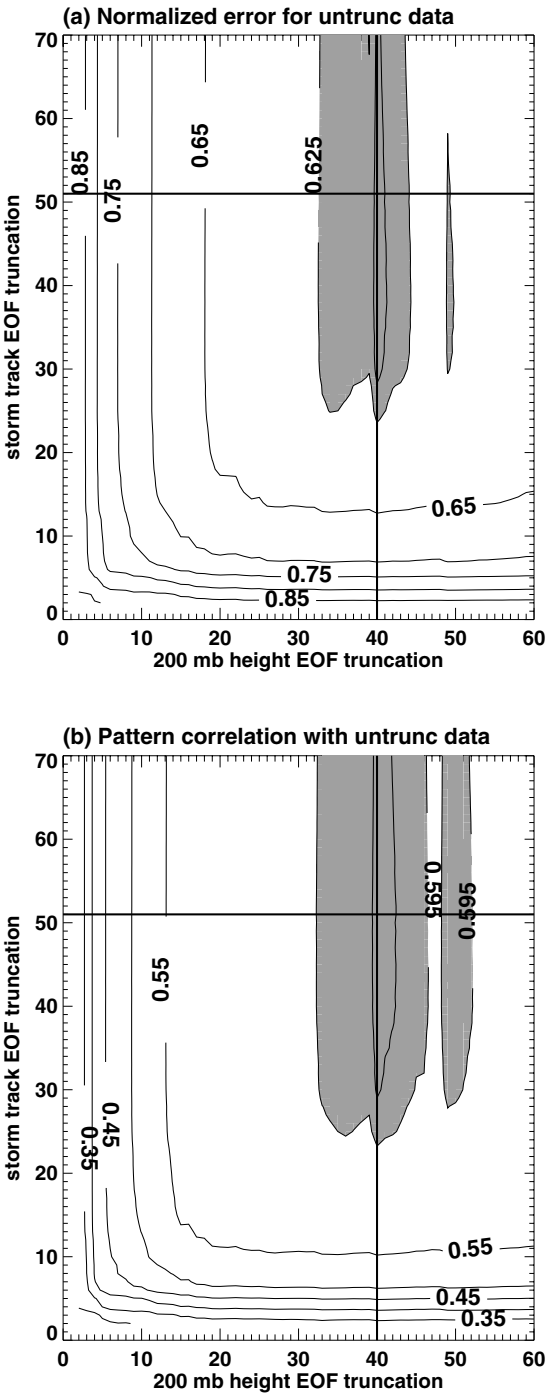

Figure 4. Cross-validated skill of the empirical storm track model as function of the number of retained EOFs of January-March 200 mb height seasonal mean and 500 mb vertical velocity bandpass variance seasonal anomalies. (a) Normalized error. (b) Anomaly pattern correlation. Contour interval is 0.05 in both panels with the addition of (a) 0.623, 0.625 and (b) 0.595, 0.598 contours. Shading begins at (a) 0.625, (b) 0.595. Horizontal and vertical lines show the truncation used for the storm track model described in the text.

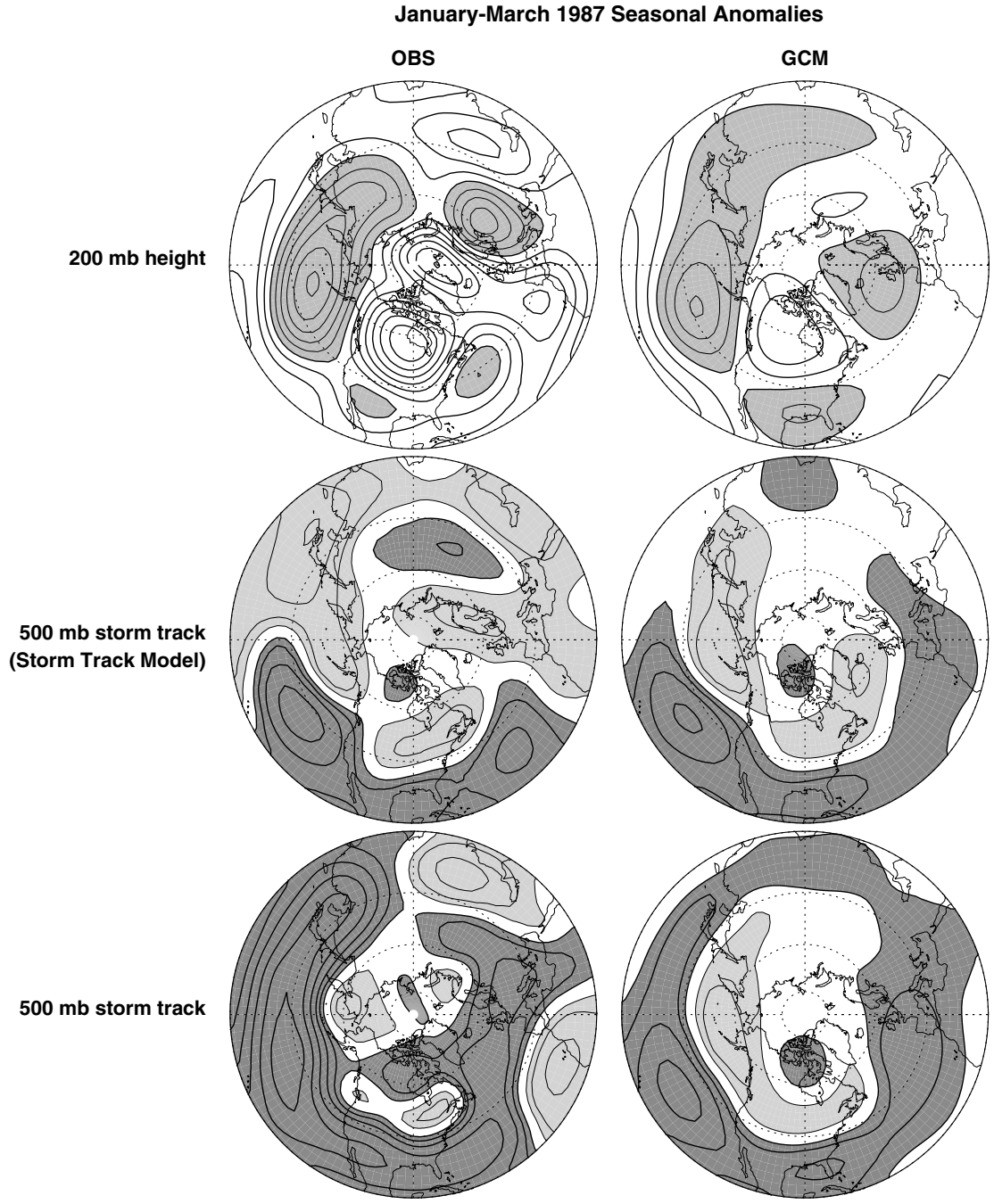

Figure 5: Seasonal anomalies for January-March 1987 from (left) NCEP reanalyses and (right) AGCM ensemble of 60 members forced with 1987 observed SSTs. (top) 200 mb height anomaly. (middle) 500 mb storm track (vertical velocity 2-7 day band passed variance anomaly) diagnosed using the empirical linear storm track model. (bottom) 500 mb storm track. The plotted quantity in the middle and bottom panels is the signed square root of the variance anomaly. Contour intervals are (top) 20m and (middle and bottom) 0.01 Pa/s with the zero contour suppressed. (top) Light shading indicates negative anomalies. (middle and bottom) Dark (light) shading indicates positive (negative) anomalies.

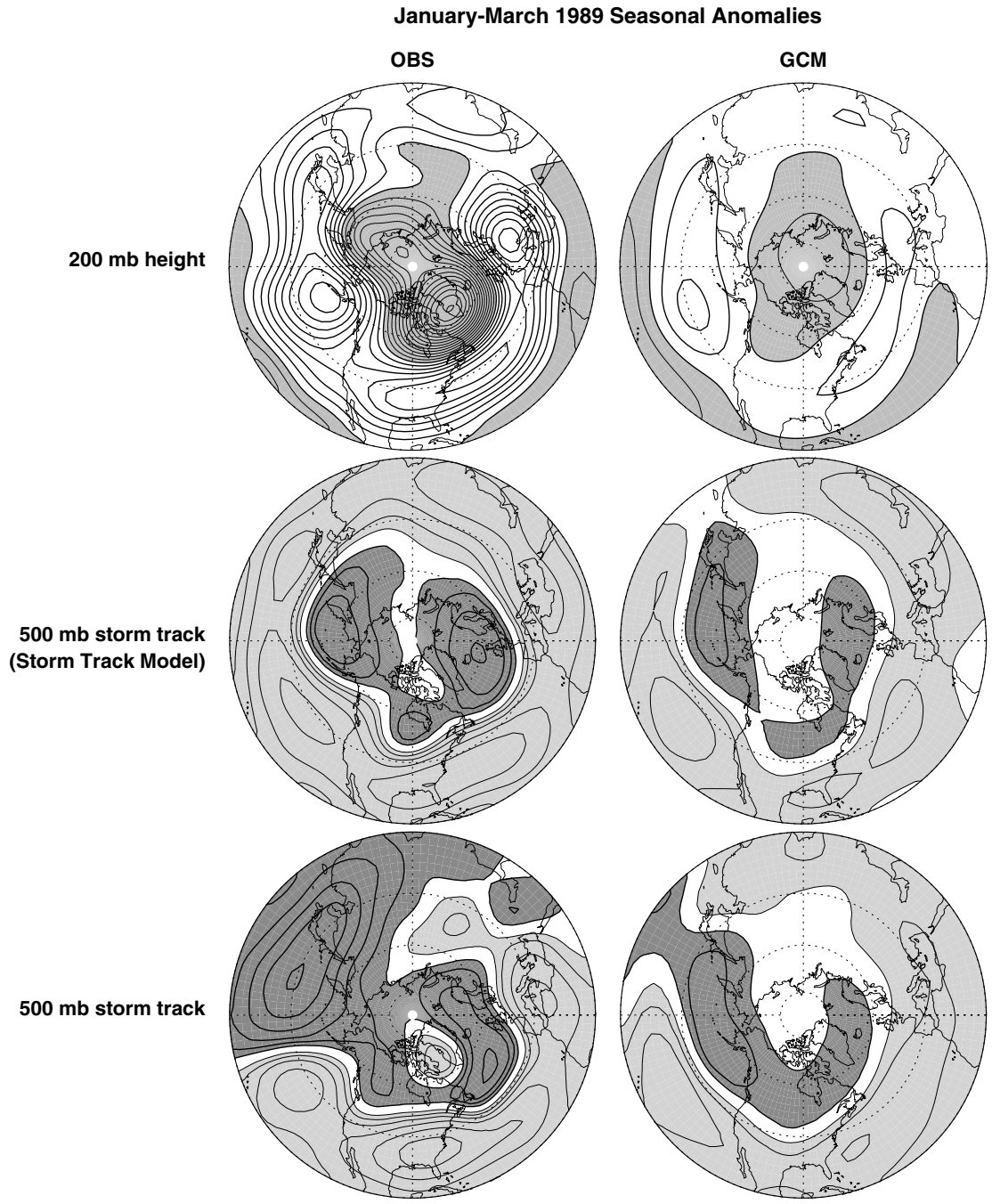

Figure 6. Same as Fig. 5 but for 1989 January-March seasonal anomalies.

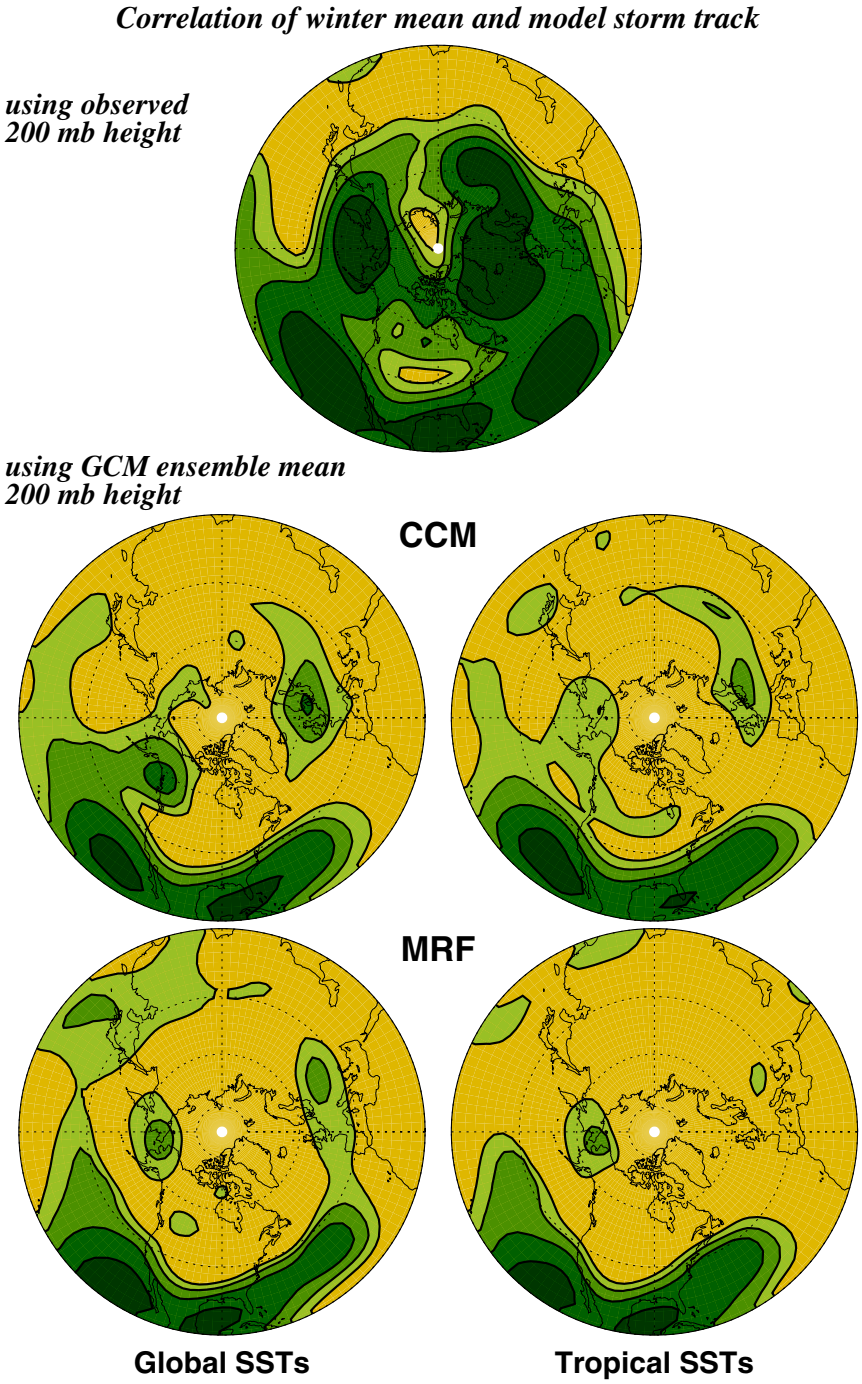

Figure 7. Predictability of storm tracks estimated using the storm track model. Green shading begins at the 5% significance level of 0.25. Contour interval is 0.15 thereafter.

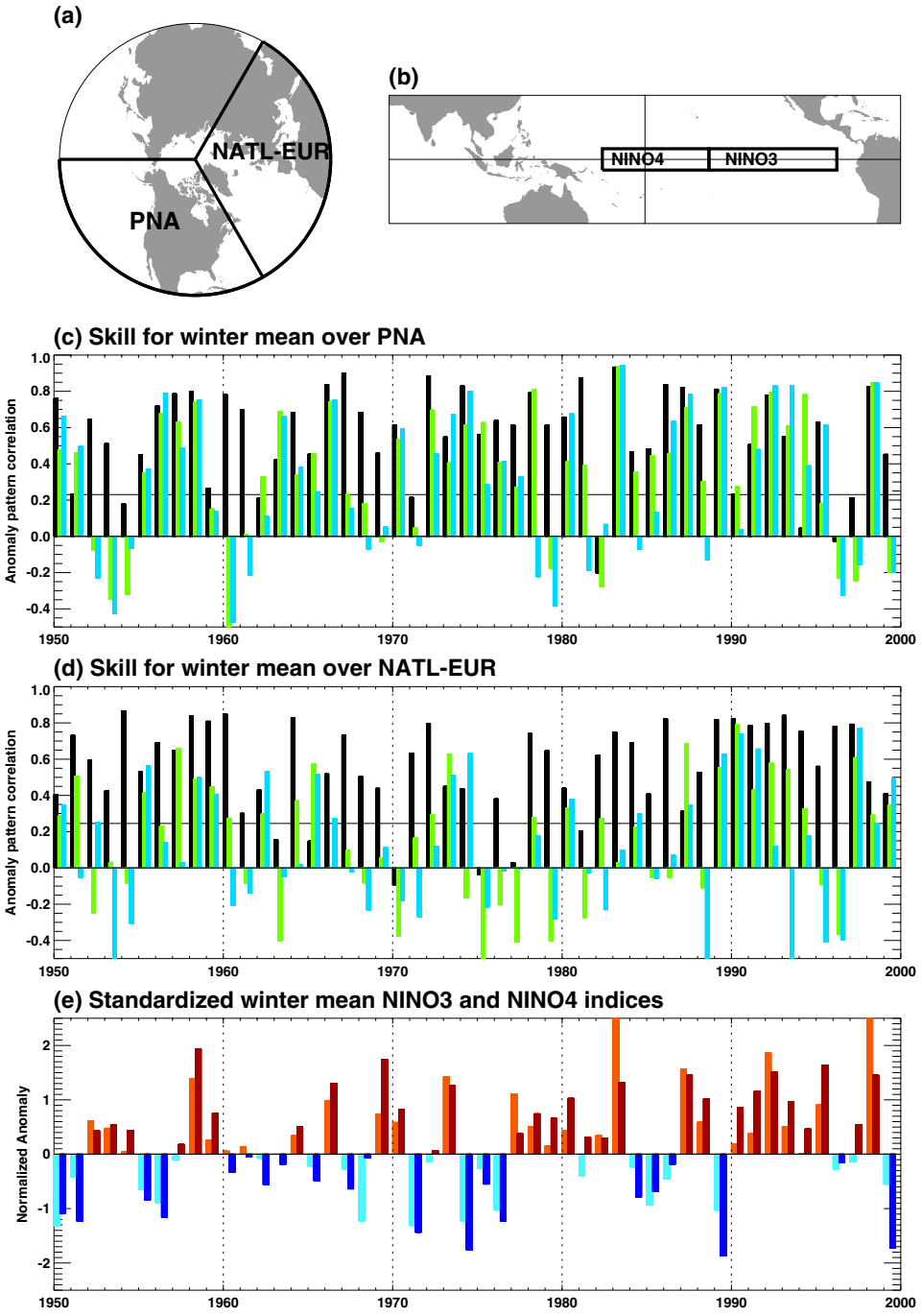

Figure 8. Timeseries of anomaly pattern correlation for January-March (JFM) storm track anomalies. (a) and (b) show regions used in subsequent panels. (c) and (d) Pattern correlation over (c) the Pacific-North American (20-90N, 180-60W) and (d) North Atlantic-European sectors (20-90N, 60W-60E) of observed and STM diagnosed storm track anomalies using 200 mb height JFM anomalies from (black bars) observed, (green bars) CCM3 AGCM forced with global SSTs, (blue bars) CCM3 AGCM forced with tropical SSTs. Thin horizontal line shows the 5% significance levels for both AGCM integrations having skill. (e) Timeseries of JFM Nino3 (orange and light blue) and Nino4 (red and dark blue) normalized by their respective standard deviations.

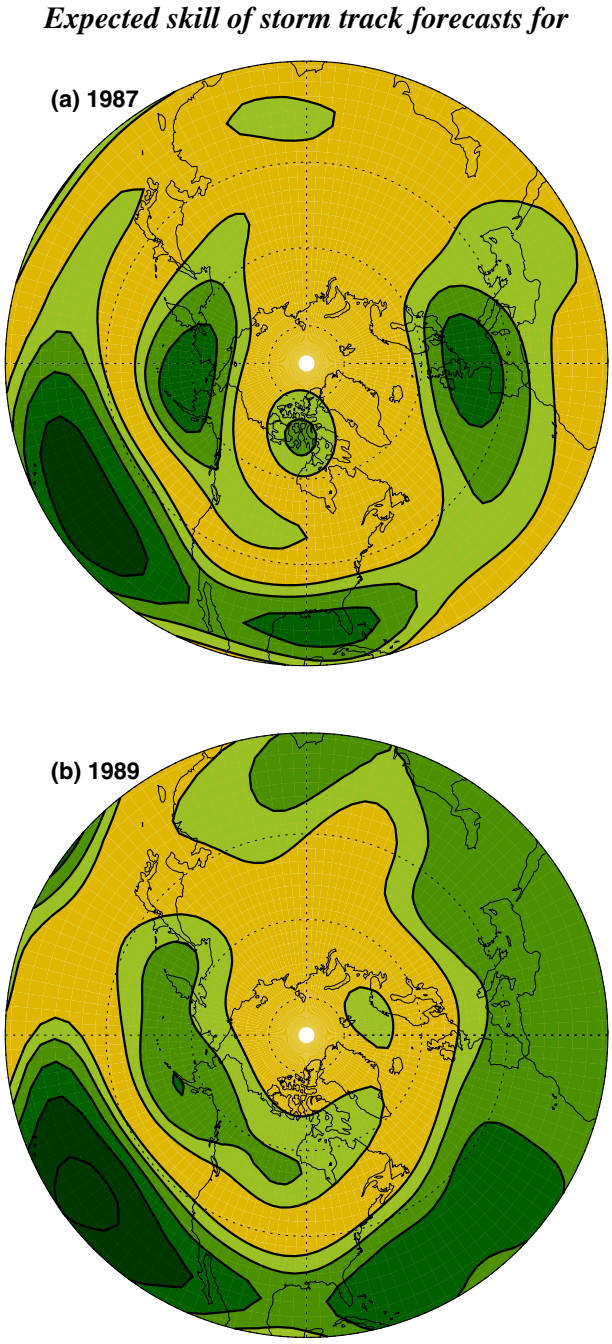

Figure 9. Case dependence of the predictability of storm tracks. Comparison of expected local anomaly correlation skill for (a) 1987 and (b) 1989 global SST forcing estimated using 60-member ensembles to calculate the signal-to-noise ratio S and then applying Eq (1) to calculate ρ_{60}. Contour interval is 0.15 starting at 0.25.

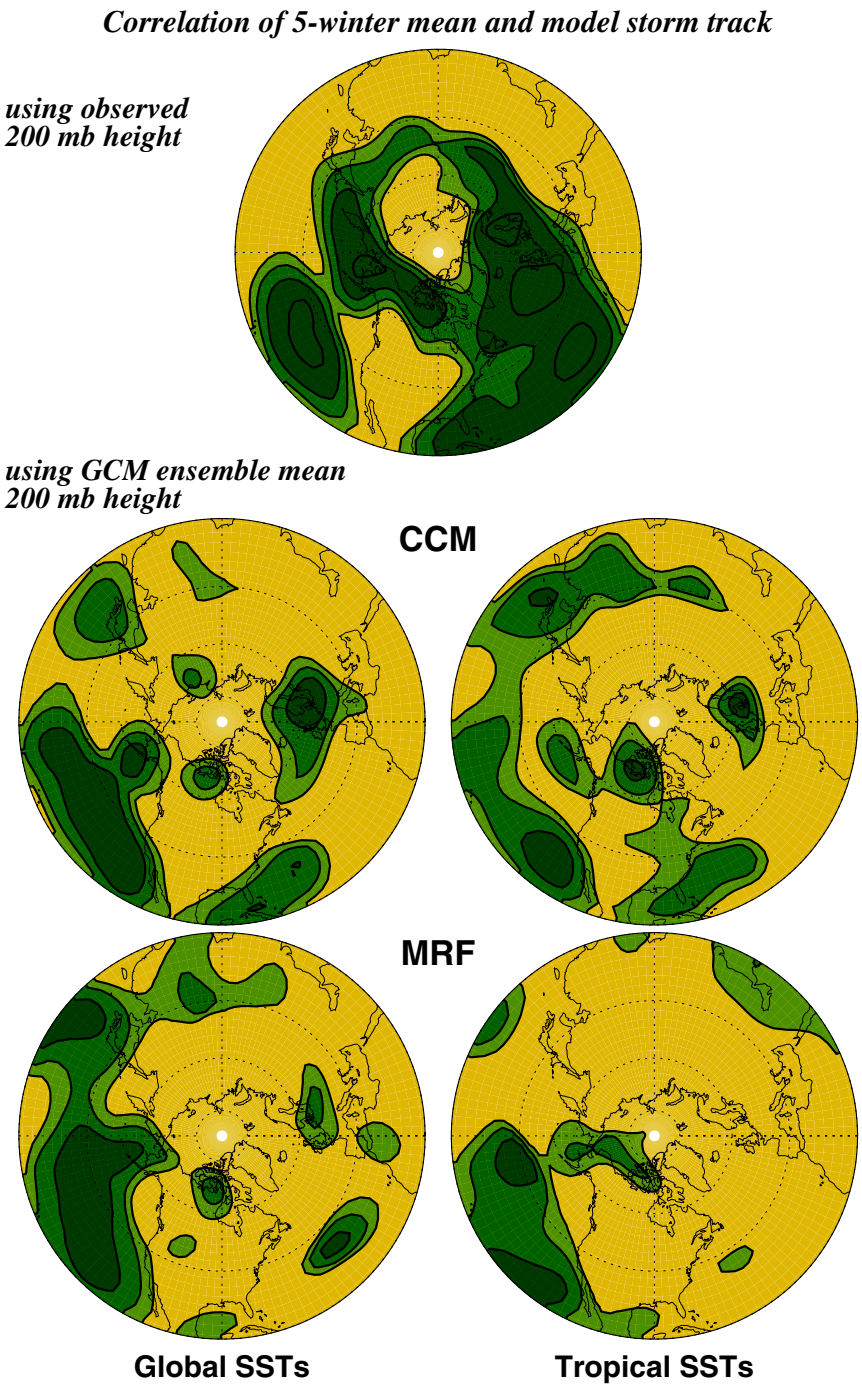

Figure 10. Skill for 5-winter averages. Green shading begins at the 5% significance level of 0.4. Contour interval is 0.15 thereafter.

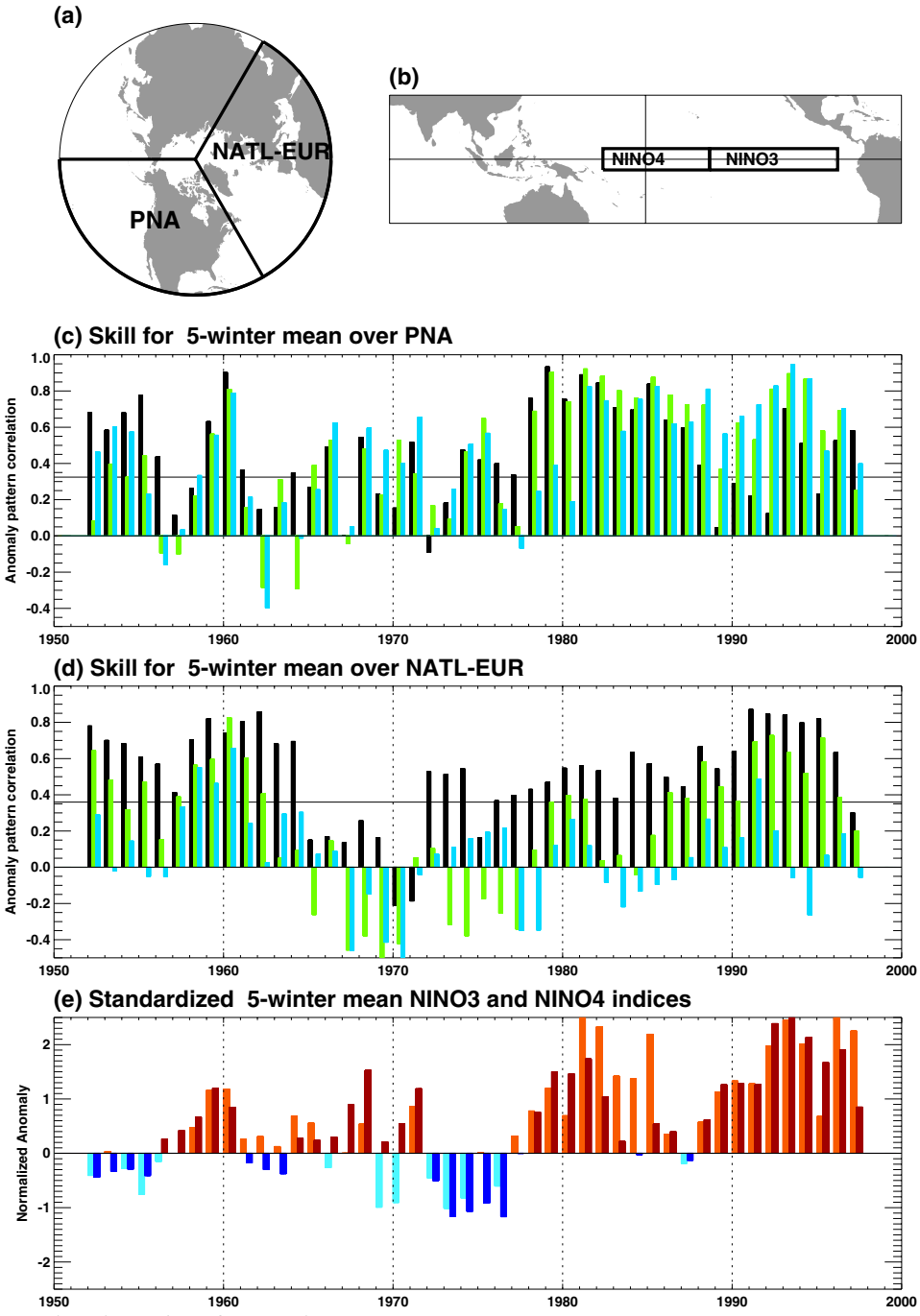

Figure 11. As in Fig. 9 but for 5-winter averages.

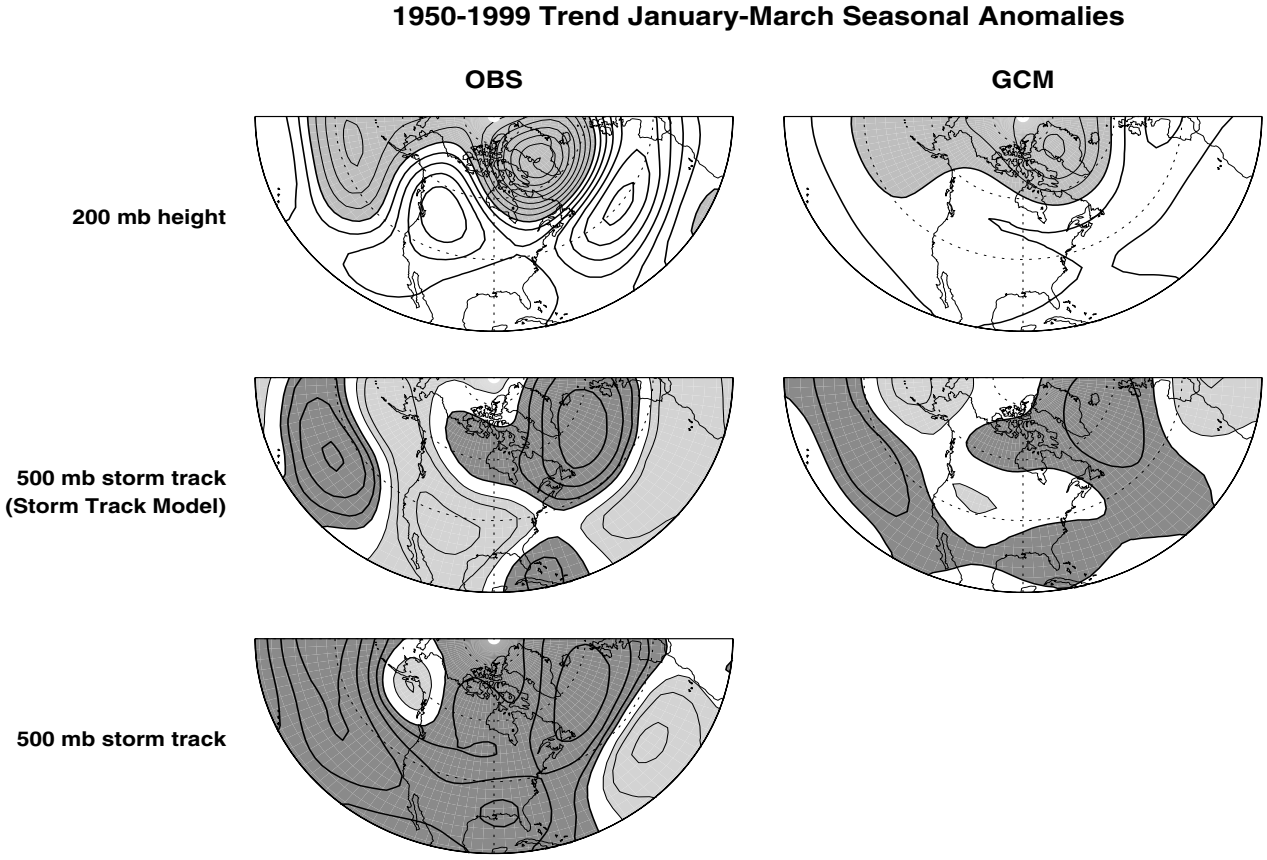

Figure 12. Linear trends for 1950-99. Plotted as the change over 50-years. (top panels) Trend in 200 mb heights from (left) NCEP-NCAR reanalysis and (right) CCM3 GOGA. Contour interval is 20m. Negative trends are shaded. (middle and bottom) Trend in the omega storm track from the STM using 200 mb heights from (left) NCEP-NCAR reanalyses and (right) CCM3 GOGA. (bottom) Trend in the stormtrack from NCEP-NCAR reanalyses. Contour intervals is 0.01 Pa s^{-1}, with the zero contour suppressed. Light (dark) regions indicate negative (positive) trends.

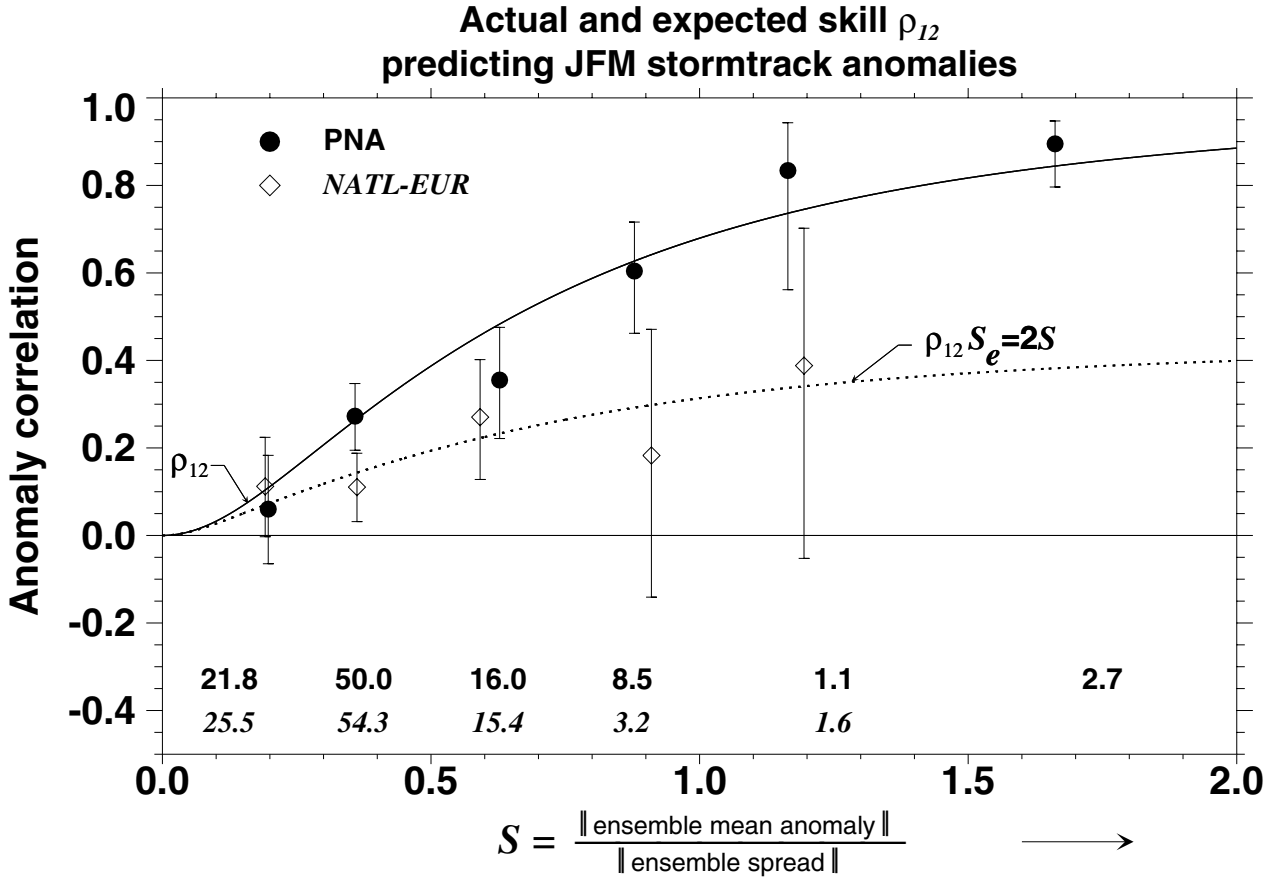

Figure 13. Anomaly correlation skill of stormtrack forecasts made using the CCM3 and MRF9 diagnosed stormtracks for January-March season. Solid curve shows the expected correlation skill ρ_n of forecasts made from the mean of n=12 member ensembles as a function of the signal to noise ratio S based on Eq (1). Dotted curve shows the expected skill ρ_{12} when a systematic error S_e=2S is present in the forecast based on Eq (4). Symbols show the actual skill of storm track forecasts for the PNA region (filled circles) and North Atlantic-European region (diamonds) binned over similar S values. Bins widths are 0.25 from S=0 to S=1 and 0.5 thereafter. Percentage of cases in each bin is indicated. Error bars show the 95% confidence interval using the Fisher z-transformation and assuming 6 esdof.